Medizinische Fakultät
der
Universität Duisburg-Essen

Aus dem Institut für Medizinische Informatik, Biometrie und Epidemiologie

Dose-Response Modeling Using Linear Splines

Inaugural Dissertation
zur
Erlangung des Doktorgrades der Naturwissenschaften in der Medizin
durch die Medizinische Fakultät
der Universität Duisburg-Essen

Vorgelegt von
Dipl. Stat. Martin Kappler
aus Berlin
2007

Dekan:	Herr Univ.-Prof. Dr. K.-H. Jöckel
1. Gutachter:	Herr Prof. Dr. M. Neuhäuser
2. Gutachter:	Frau Priv.-Doz. Dr. U. Krämer, Düsseldorf
Tag der mündlichen Prüfung:	3. Dezember 2007

A part of the present work has been published or has been submitted in the following journals:

Kappler, M., Pesch, B., Marczynski, B., Rihs, H.P., Angerer, J., Scherenberg, M., Adams, A., Seidel, A., Wilhelm, M. and Brüning T. (2006): Dose-response relation of external and internal exposure and genotoxic effects in workers with exposure to PAH. Proceedings of the 46th Meeting of the German Society for Occupational and Environmental Medicine (DGAUM), Arbeitsmedizin, Sozialmedizin, Umweltmedizin, 41, 108. [*german*]

Pesch*, B., Kappler*, M., Straif, K., Marczynski, B., Preuss, R., Roßbach, B., Rihs, H.P., Weiss, T., Rabstein, S., Pierl, C., Scherenberg, M., Adams, A., Käfferlein, H.U., Angerer, J., Wilhelm, M., Seidel, A. and Brüning, T. (2007): Dose-response modeling of occupational exposure to polycyclic aromatic hydrocarbons with biomarkers of exposure and effect. Cancer Epidem. Biomar. Prev., 16, 1863-1873.
*equally contributed

All models are wrong, but some are useful.

George E. P. Box

Contents

1 Introduction

Modeling an unknown relation between an outcome and a continuous predictor is a common problem not only in epidemiology. It is carried out to describe and understand the mechanism of the dependencies between the variables in question. In occupational medicine, statistical modeling of the associations between external and internal dose as well as dose-response relations between exposure and effect in biological monitoring has become important for risk assessment and in the regulatory process for chemicals of concern. Knowledge about the strength of the association and the shape of the dose-response function can support a health-based setting of limits of exposure.

An easy starting point for statistical analysis is given by assuming a linear relation between the studied variables. However, many reasons can be given as to why the assumption of linearity may be transgressed, e. g. because of threshold effects, sensitization, saturation, or in occupational settings, due to the existence of a *healthy worker effect*. A common alternative to the assumption of linearity is to categorize the exposure measure and to assume constant levels of the outcome within each category. The advantages of this approach are an easy interpretation and communication of the results using tables. However, the loss of information on an often carefully investigated exposure measure can not be neglected. This approach is accompanied by important drawbacks and is criticized by many authors (Greenland, 1995b; Thurston, Eisen and Schwartz, 2002; Steenland and Deddens, 2004; Altman and Royston, 2006). One of the criticisms is that the assumption of constant outcome levels within the categories seems unrealistic in most cases. Selection of the category boundaries is often conducted somewhat arbitrarily while it can have a large influence on the estimated shape of the dose-response curve. Furthermore, misclassification in the categories due to measurement error of the predictor variable may introduce differential misclassification into the analysis. Simulation studies showed a loss of efficiency by categorizing quantitative exposure measure (Zhao and Kolonel, 1992; Greenland, 1995b).

Numerous other statistical methods are investigated for modeling unknown functional relations to cope with these problems, especially in the field of non-linear or non-parametric regression. On the one hand, these methods offer the advantage of fitting unknown functional relations in a flexible way. On the other hand, model parameters are either difficult to interpret or not present at all for these methods. However, for the application of estimators and models in regulatory processes, model parameters have to be easy to communicate and

have to ensure biological plausibility. Toxicological findings suggest, for example, thresholds for some substances. Furthermore, in the low-dose range, effects are described as showing deviations of the general noxious effects also known as *hormesis*. In the high-dose range, some sort of *overload* phenomena are discussed that initiate further pathomechanisms. Therefore, the dose-response curve might exhibit quite a complex shape.

Examples of methods capable of detecting such complex curves are fractional polynomials (Royston and Altman, 1994), additive models (Hastie and Tibshirani, 1986) and linear splines (de Boor, 1978). The main focus of this work is on the latter that exhibit some advantages for modeling of a continuous predictor. The model's features are a compromise between sufficient flexibility to reproduce the underlying relation, simplicity of graphical illustrations and easy interpretability of the model parameters. Splines are piecewise polynomials that join in a smooth way at the edges, the so-called *knots*. The feature of smoothness degenerates for linear splines to continuity, comparable to a frequency polygon. Thus, in each category a linear relation between exposure and outcome is modeled. The categories can be chosen either in a user-driven manner or by the use of adequate criteria for an *optimal* assignment.

All three methods are presented in detail within this work and are compared to the standard techniques for modeling dose-response relations. The methods are applied on an occupational dataset of workers exposed to polycyclic aromatic hydrocarbons (PAHs) conducted at the Research Institute of Occupational Medicine of the German Social Accident Insurance (BGFA, Bochum, Germany) from 1999 to 2004 in different occupational settings in Germany.

Chapter 2 gives an overview of the material and methods used within this work. Section 2.1 introduces the PAH study; the conduct of the study is outlined and the techniques applied for the assessment of occupational PAH exposure and for the measurement of urinary metabolites are described. The statistical methods that are linear splines, fractional polynomials, additive models, analysis of variance (ANOVA) and linear regression, used for the analyses of the dose-response relation between occupational exposure to phenanthrene (PHE) and urinary excretion of the sum of 1-, 2-+9-, 3- and 4-OH-phenanthrene (OHPHE), are described in detail in the sections 2.2 to 2.5. Section 2.6 introduces methods for a comparison of the fit of different models.

The results of the statistical analyses are presented in chapter 3. Section 3.1 gives an overview of the distribution of demographic variables as well as variables of external and internal exposure to PAHs. Results of the dose-response analyses using the different methods are detailed in section 3.2. For each method, parameter estimates are presented, the curve of the dose-response relation is illustrated and the model fit is evaluated. The section concludes with a comparison of the model fit between the applied methods.

Chapter 4 critically discusses the methods applied for analysis of the dose-response relation. The main areas of application as well as benefits and drawbacks of each method are illustrated. The model selection procedures are compared to recommendations in the literature. Particularly, the model selection approach and the restrictions for the choice of the final linear splines model are examined in detail.

Finally, in chapter 5 the author's conclusions are outlined.

2 Material and Methods

2.1 Polycyclic Aromatic Hydrocarbon Study

2.1.1 Background

Polycyclic aromatic hydrocarbons (PAHs) represent complex mixtures composed of carbon and hydrogen atoms fused as benzenoid rings or as unsaturated four- to six-membered rings (Rihs et al., 2005). PAHs are generally formed during pyrolysis and incomplete combustion of organic matter in a variety of occupational and environmental settings (Bostrom et al., 2002; Hemminki et al., 1990). Occupational exposure to PAHs occurs during coke production, manufacture of refractory products and graphite electrodes, in foundries and many other processes (IARC, 1984; Straif et al., 2005). PAHs are established lung carcinogens (Doll et al., 1972; IARC, 1983).

The composition of PAHs varies by many factors. Among the more than 500 PAHs detected, there has been no common international agreement on which compounds should be reported concerning human exposure in environmental or occupational settings (Bostrom et al., 2002). The choice of indicator substances has resulted from historical, toxicological and other considerations. In particular, pyrene and phenanthrene (PHE) are abundant PAHs which are frequently measured as indicators of external exposure. Their metabolites 1-OH-pyrene and the different OH-phenanthrenes serve as biomarkers for PAH exposure in humans.

PAHs serve as an example of an environmental and occupational group of chemicals with extensive biomonitoring research. Monohydroxylated PAHs have been employed as biomarkers for human exposure assessment.

2.1.2 Conduct of the Study

In order to characterize the dose-response relationship between external and internal exposure and to evaluate potential genotoxic effects of PAHs, a cross-sectional biomonitoring study in German workers was conducted in different occupational settings with PAH exposure in collaboration between academia and industry. Results regarding the impact of a technological improvement on biomarkers of exposure, as well as results about the impact of exposure

and genetic polymorphisms on the urinary metabolite concentrations have been published (Marczynski et al., 2005; Rihs et al., 2005). The genotoxic effects were evaluated for coke-oven and graphite-electrode-producing workers (Marczynski et al., 2002).

The study data was collected between March 1999 and May 2004. The study population comprised workers occupationally exposed to PAH during manufacture of refractory products and graphite electrodes, coke oven works, tar distillation, and infeed of converters. To measure the exposure to PAHs, each worker carried a personal air sampler during the shift. After the shift, the workers provided a sample of spot urine to measure the internal exposure and blood samples to assess genotoxicity. A structured questionnaire was applied in a face-to-face interview to assess demographic characteristics and smoking habits, among other data. All study subjects provided a written informed consent prior to investigation. The study was approved by the ethic commission of the Ruhr University Bochum and conducted in accordance with the definitions of the Declaration of Helsinki (WMA, 1964).

2.1.3 Determination of Occupational Exposure to Airborne PAHs

Personal air sampling was conducted in the worker's breathing zone for an average of two hours to assess exposure to PAHs for a working shift. Samples were collected with battery-operated personal air sampling pumps according to a procedure of the German Institute for Occupational Safety and Health (BGIA, Sankt Augustin, Germany). Sixteen U.S. EPA PAHs were analyzed according to method 5506 published by the U.S. National Institute for Occupational Safety and Health (NIOSH, 1994). The limit of quantification (LQ) ranged between 0.007-0.51 $\mu g/m^3$ for the different PAHs. Observations below the LQ were set to half of the LQ.

2.1.4 Determination of PAH Metabolites, Cotinine, and Creatinine in Urine

Spot urine samples were collected at the end of the work shift in polypropylene tubes and frozen at $-20\,^{\circ}C$ until preparation. The determination of 1-OH-pyrene and of 1-, 2-+9-, 3- and 4-OH-phenanthrene was carried out by high performance liquid chromatography (HPLC) with fluorescence detection, as described in Lintelmann and Angerer (1999) and Marczynski et al. (2002). Briefly, the metabolites were enriched on a pre-column, consisting of copper phthalocyanine-modified silica gel, separated on a RP-C18 column and quantified by fluorescence detection. The LQ ranged between 24 and 96 ng/L for the different metabolites. Urinary creatinine (crn) was determined photometrically as picrate, according to the Jaffé method (Taussky, 1954). The concentration of the metabolites was presented in $\mu g/g\,crn$. Urinary cotinine was determined by gas chromatography with nitrogen-specific detection after a liquid/liquid extraction of the urine samples, according to the procedure described by Scherer et al. (2001).

2.2 Splines

Splines are piecewise polynomials that join in a smooth way at the edges and are a flexible way to fit an unknown functional form (de Boor, 1978). They are a wide class of functions that are used in many areas such as computer graphics and the design of cars and aircrafts. The term *spline* was introduced by Schoenberg (1946) and has its origins in shipbuilding. Wooden slats that are fixed with nails take the form of a cubic spline.

A variety of spline applications exist in the field of statistics especially for linear, non-linear and non-parametric regression such as regression splines and smoothing splines, additive models (see section 2.4) and multivariate adaptive regression splines. In this section, regression splines will be presented. These are also known as least squares splines.

The formal definition of a spline is a function $S : [a, b] \to \mathbb{R}$ consisting of polynomial pieces $P_i : [t_{i-1}, t_i] \to \mathbb{R}$ with $a = t_0 < t_1 < \ldots < t_k < t_{k+1} = b$ and $S(x) = P_i(x) \; \forall \; t_{i-1} \leq x \leq t_i$. The points t_i, $i = 1, \ldots, k$ are called interior knots whereas t_0 and t_{k+1} are called exterior knots of the spline. If each piecewise polynomial is of degree d, the spline is said to be of degree d. To ensure smoothness, the spline is requested to have $d - 1$ existing derivatives at each of k knots $t_1 \ldots, t_k$.[1] In mathematical terms this means

$$\lim_{x<t_i} S^{(j)}(x) = \lim_{x>t_i} S^{(j)}(x) \quad \forall i = 1, \ldots, k \text{ and } j = 1, \ldots, d-1 \; .$$

2.2.1 Basis of the Space of Splines

The family of splines of a given degree d and with given knots $t_1, \ldots, t_k$ form a linear space of functions (de Boor, 1978). A spline of this family can be constructed by linear combinations of basis functions.

The choice of the basis for the space of splines is arbitrary, as all regression parameters from the use of one basis can be transformed into those from another basis. In this section, the truncated power basis and the B-spline basis are presented. However, for the application of linear splines, these bases have some disadvantages. Therefore, a new basis for the space of linear splines is developed in the last paragraph. This new basis avoids some of the drawbacks and will therefore be used for the application of linear splines in this work.

[1] If the existence of the d^{th} derivative would be claimed as well, the spline would degenerate to a polynomial of degree d on the whole domain and would loose much of its flexibility.

2.2.1.1 Truncated Power Basis

The most intuitive basis for the linear space of splines is the truncated power basis, sometimes also referred to as the natural basis. It consists of the following functions (de Boor, 1978):

$$1,\ (x-t_0),\ (x-t_0)^2, \ldots,\ (x-t_0)^d,\ (x-t_1)_+^d, \ldots,\ (x-t_k)_+^d\ .$$

with $(z)_+ = \max(0, z)$. Sometimes, the first $d+1$ functions of this basis are defined slightly differently as $1,\ x,\ x^2, \ldots,\ x^d$. However, this only changes the interpretation of the parameters.

The last k basis functions are each linked to one different interior knot, which can be advantageous from a modeling point of view (Hansen and Kooperberg, 2002), e. g. for variable selection methods. However, the truncated power basis exhibits rather poor numerical properties. In regression problems the design matrix deteriorates and may no longer be invertible with an increasing number of knots.

Figure 2.1: Truncated power basis $\mathrm{TP}_{i,t,d}$ for linear splines ($d = 1$) with two interior knots $t = (0.4, 1.4)'$

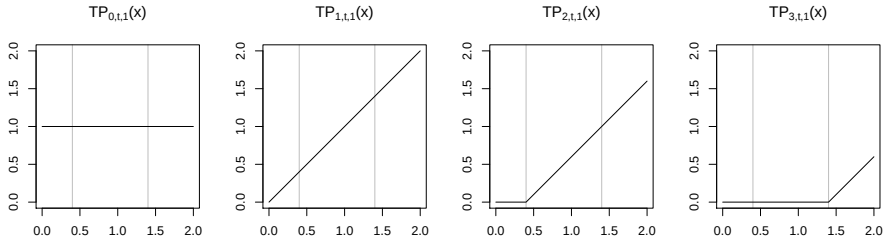
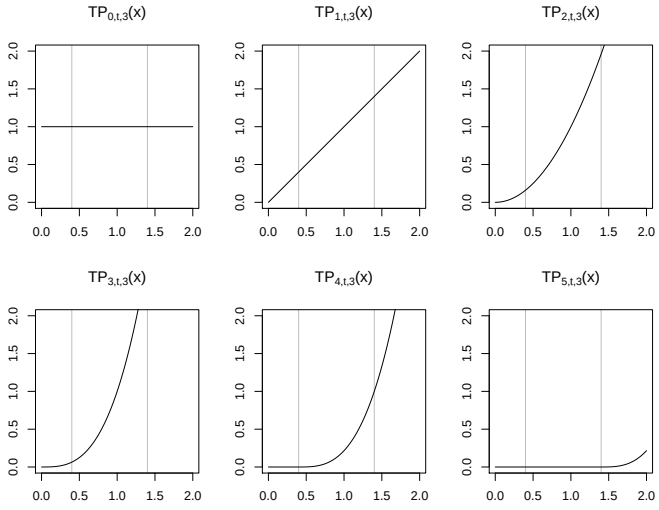


Figure 2.2: Truncated power basis $\mathrm{TP}_{i,t,d}$ for cubic splines ($d = 3$) with two interior knots $t = (0.4, 1.4)'$



Figures 2.1 and 2.2 show the truncated power basis for splines of degree $d = 1$ and of degree $d = 3$, each with two interior knots.

2.2.1.2 B-Splines

The B-splines[1] were first developed by Schoenberg (1946). However, only after the work of de Boor (1978) were they widely accepted and applied. The formal definition of B-splines can be found in either of these works, but will not be given in detail here.

The great advantage of B-splines over the truncated power basis is its stable computation. Because of the minimum support property[2], the design matrix shows a block diagonal form and is easier to invert. Examples of the linear and cubic B-spline functions for splines of degree 1 and 3, each with two interior knots, are shown in Figures 2.3 and 2.4.

Figure 2.3: B-spline basis $B_{i,t,d}$ for linear splines ($d = 1$) with two interior knots $t = (0.4, 1.4)'$

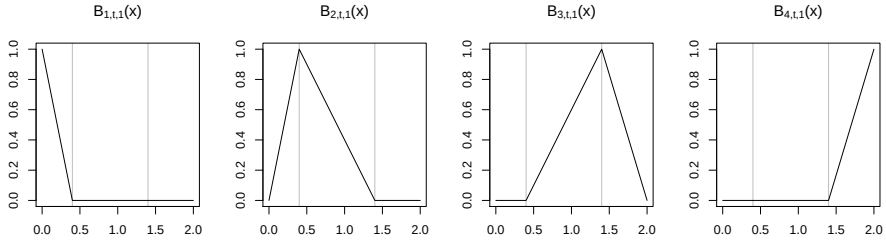
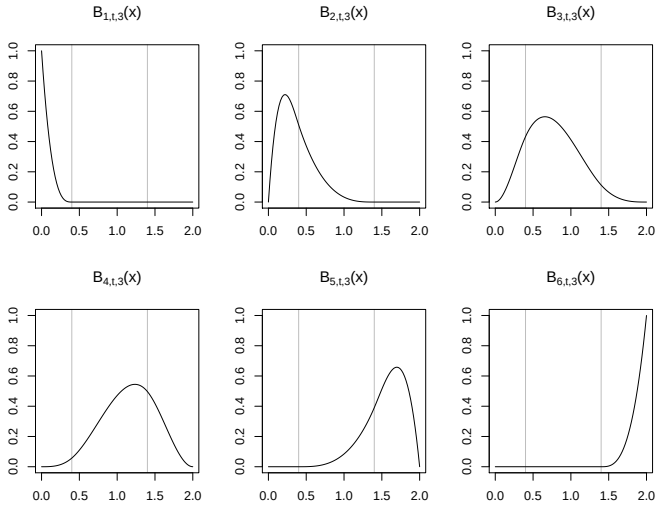


Figure 2.4: B-spline basis $B_{i,t,d}$ for cubic splines ($d = 3$) with two interior knots $t = (0.4, 1.4)'$



[1] B stands for *basis*.
[2] That means, B-splines are the basis with the smallest region of function values $\neq 0$ for each basis function (de Boor, 1978).

2.2.1.3 Double Truncated Linear Basis

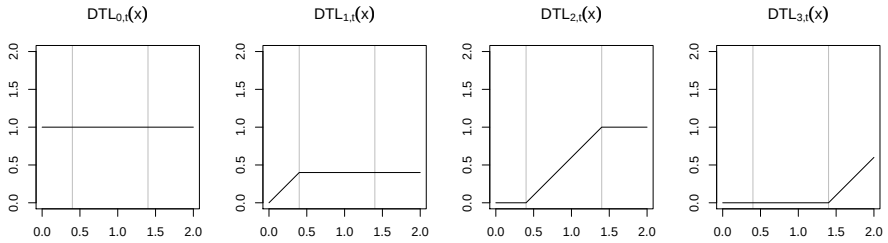
A new basis for the space of linear splines is presented in this paragraph. Due to its construction it is denoted as the double truncated linear basis (DTL basis). The DTL basis is defined by the following formulae:

$$
\begin{aligned}
\mathrm{DTL}_{0,t}(x) &= 1 \\
\mathrm{DTL}_{1,t}(x) &= \begin{cases} x - t_0 & \text{if } x < t_1 \\ t_1 - t_0 & \text{if } x \geq t_1 \end{cases} \\
&= I_{(-\infty,\, t_1)} \cdot (x - t_0) + I_{[t_1,\, \infty)} \cdot (t_1 - t_0) \\
\mathrm{DTL}_{i,t}(x) &= \begin{cases} 0 & \text{if } x < t_{i-1} \\ x - t_{i-1} & \text{if } t_{i-1} \leq x < t_i \\ t_i - t_{i-1} & \text{if } x \geq t_i \end{cases} \\
&= I_{[t_{i-1},\, t_i)} \cdot (x - t_{i-1}) + I_{[t_i,\, \infty)} \cdot (t_i - t_{i-1}) \quad \forall\, i = 2, \ldots, k \\
\mathrm{DTL}_{k+1,t}(x) &= \begin{cases} 0 & \text{if } x < t_k \\ x - t_k & \text{if } x \geq t_k \end{cases} \\
&= I_{[t_k,\, \infty)} \cdot (x - t_k)
\end{aligned}
$$

with $t = (t_1, \ldots, t_k)$ the interior knots, t_0 and t_{k+1} the exterior knots with $t_0 < t_1 < \ldots < t_k < t_{k+1}$, and $I_{[z_1,\, z_2]}$ the indicator function for the interval $[z_1,\, z_2]$.

An example of this basis for two interior knots is given in Figure 2.5.

Figure 2.5: Double truncated linear basis $\mathrm{DTL}_{i,t}$ with two interior knots $t = (0.4, 1.4)'$



The proof that the DTL functions defined above form a basis of the space of linear splines is given in Appendix A.

The application of the DTL basis in regression leads to the following model

$$
y = \sum_{i=0}^{k+1} \beta_i \cdot \mathrm{DTL}_{i,t}(x) + \varepsilon
$$

with ε being the error term following a normal distribution with stochastic independence between the observations, i. e. $\varepsilon \sim N(0, \sigma^2)$.

For each two adjacent knots t_{i-1}, t_i there is only one corresponding DTL function, DTL_i, that has a positive slope between these two knots. As this slope is equal to unity, the corresponding regression parameter β_i represents the slope of the regression spline in the segment (t_{i-1}, t_i) $(i = 1, \ldots, k+1)$. Estimates of β_i can be interpreted directly as estimates of this slope. Tests for a significant deviation of a zero slope can be constructed by testing the hypothesis $H_0: \beta_i = 0$, while the difference of slopes in two consecutive segments can be tested by means of the hypothesis $H_0: \beta_{i-1} = \beta_i$.

2.2.2 Regression Splines

Regression splines are the application of splines in the field of linear and non-linear regression. That means, the usual regression model $y = \beta_0 + \beta_1 x + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$ is extended to

$$y = \sum_i \beta_i B_{i,t,d}(x) + \varepsilon$$

with y the vector of the outcome, x the vector of the predictor, $B_{i,t,d}(x)$ the i^{th} function of a spline basis. As the basis functions depend on the knots $t = (t_1, \ldots, t_k)$ and the degree d of the piecewise polynomials, questions of how to determine these quantities need to be addressed.

In the application of regression splines, cubic splines are generally chosen, i. e. polynomials of degree $d = 3$ between two subsequent knots (Brown, Ibrahim and DeGruttola, 2005). Another possibility is the choice of the degree $d = 1$, i. e. the use of linear splines. These are piecewise linear functions that are connected at the knots (Molinari, Daures and Durand, 2001; Rosenberg et al., 2003). This leads to a simple parametric model as recommended by Steenland and Deddens (2004), because a knot could be interpreted as threshold where a function alters its characteristic.

2.2.3 Knot Selection

In order to determine the knot locations, the knots are added as regression parameters. This changes the linear into a non-linear model. However, the only real non-linear parameters in the model are the knots, i. e. given the knots, estimates of the other parameters can be obtained by usual linear regression methods. This reduces the minimization algorithm to just the knot parameter space resulting in much less computational effort. Such a model is denoted *separable* and the parameters that can be obtained by linear regression are called *conditionally linear* (Smyth, 2002).

In this work, the applied procedure was as follows. The residual sum of squares (RSS) was calculated as a function of the knots only and minimized using the Levenberg-Marquardt algorithm[1] (Moré, 1978). The optimization procedure was performed on a grid of about 200 initial points in the space of the knots to reduce the possibility of ending in a local minimum.

In order to avoid over-fitting, models were considered only if each segment contained at least 10% of the observations between two consecutive knots. As a consequence, knot positions could only be determined between the 10^{th} and 90^{th} percentile of the predictor.

To test whether the inclusion of an additional knot results in a better model fit, the following statistic can be applied

$$F = \frac{(RSS_0 - RSS_1)/r}{RSS_1/(n-p)}\,, \tag{2.1}$$

with RSS_0 the residual sum of squares of the smaller model, RSS_1 that of the model with additional parameters, r the difference in the number of parameters of the two models, n the number of observations, and p the number of parameters of the larger model. Under the null hypothesis of no necessity of the larger model, F asymptotically follows a F-distribution with r and $n-p$ degrees of freedom (Smyth, 2002). In the comparison of a model with one additional knot against the model without this knot, the number of parameters differs by 2, the knot itself and the regression parameter β_i of the i^{th} segment, i. e. $r = 2$. In order to test whether the inclusion of the respective knot results in a better model fit, the statistic F of (2.1) is compared to the $(1-\alpha)$-quantile of the F-distribution with 2 and $n-p$ degrees of freedom. If $F > F_{2,n-p,1-\alpha}$, the model with the specified knot is preferred over that without the knot.

In order to choose the best fit model, the number of knots is increased one at a time and the described F-test with the statistic of (2.1) is conducted. However, the selection procedure should not be stopped immediately after one non-significant test. Some data situations may be such that the inclusion of one knot does not contribute to a better model fit, whereas the inclusion of a second or third does. A straightforward rule, for how many further knots have to be considered after a non-significant test, cannot be given. Nonetheless, without discontinuities in the underlying functional relationship, looking ahead two or three steps should generally suffice.

2.3 Fractional Polynomials

Fractional polynomials have been introduced by Royston and Altman (1994) as an alternative to traditional approaches for the analysis of continuous outcomes. They are a "sensible com-

[1] as implemented in SAS/IML (SAS Institute Inc., Cary, NC, USA)

promise between really complex curves and over-simplified straight lines" (Royston, Ambler and Sauerbrei, 1999).

The starting point of fractional polynomials is the simple linear regression with the straight line $\beta_0 + \beta_1 x$. In some cases, this already might be an adequate description of the underlying functional relationship. However, fractional polynomials introduce additional power transformations x^p of the continuous predictor x as variables into the model. Royston and Altman propose to restrict p to only a set of 8 values $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where $p = 0$ denotes the transformation with the natural logarithm $\ln(x)$. For example, if $p = 0.5$ the adjusted function would be $\beta_0 + \beta_1\sqrt{x}$. A function with p chosen from this set is called a fractional polynomial of order $m = 1$. The set of powers has not been changed since the original work from Royston and Altman. Even if the set is small, it provides considerable flexibility.

When applying the transformations of the set S to x, it is important to note, that some of the resulting functions are only defined for positive values, notably for $p \in \{-0.5,\ 0,\ 0.5\}$. All values of x have to be positive. If this is not the case, Royston and Altman propose to add a constant to the values of x.

In order to decide which model to choose, all power transformations are individually fit to the data by usual linear regression and the corresponding model deviances are calculated. The deviance D_M of a model M is defined as two times the difference of the log-likelihood of M and that of a saturated model, i. e. a model that has as many parameters as observations (McCullagh and Nelder, 1989)

$$D_M = -2(\ln L_M - \ln L_{\text{saturated}}) \ .$$

The maximum difference between the deviances of all models with $p \in S\backslash\{1\}$ and that of the linear model $p = 1$ is approximately χ^2-distributed with 1 degree of freedom. If this difference is greater than the 95^{th}-quantile of the χ^2_1-distribution, the corresponding model is preferred over the straight line. The saturated model does not have to be specified because only deviance differences are considered and the corresponding term drops out.

Higher order fractional polynomials can be used to achieve a considerably higher amount of flexibility. In this way, many functions with one single minimum or maximum such as J-shaped relationships can be modeled effectively. Generally, $m \leq 2$ is chosen for every continuous predictor in the model. The resulting model for a fractional polynomial of order $m = 2$ is of the form $\beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$ with $p_1,\ p_2 \in S$. For $p_1 = p_2 = p$, the following model is considered $\beta_0 + \beta_1 x^p + \beta_2 x^p \ln(x)$. All possible models ($N = 36$) are fit to the data and the model deviances are calculated. The maximum difference in deviances between all second order models and the best fit first order model approximately follows a χ^2-distribution with 2 degrees of freedom. If this maximum difference is greater than the 95^{th}-quantile of the

χ^2_2-distribution, the corresponding model is preferred over the best fit first order model.

2.4 Additive Models

Additive models are a non-parametric or semi-parametric regression technique. They were first introduced in the early eighties mainly via the work of Stone (1985) and subsequently promoted via the work of Hastie and Tibshirani (1986), Buja, Hastie and Tibshirani (1989) as well as Hastie and Tibshirani (1990) under the more general viewpoint of generalized additive models[1]. In contrast to non-parametric regression, the independent variable is modeled by a univariate smoother. This smoother models the unknown functional form of the relationship between outcome and the independent variable.

Additive models are a generalization of the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ by replacing the $\beta_1 x$ with some smooth non-parametric functions, i. e.

$$y = \beta_0 + s_1(x) + \varepsilon \ .$$

A logical extension of the model is to include a univariate smoother for each independent continuous variable. However, as the intention of this work is to describe the shape of the dose-response curve between one predictor and the outcome, notation is restricted to one predictor.

Further predictors Z_i that are assumed to have linear effects on the outcome can be considered in the model as well. This leads to a semi-parametric model of the form

$$y = \beta_0 + s_1(x) + \sum_{i=1}^{k} \beta_i Z_i + \varepsilon \ .$$

In order to ensure estimability, the function s_1 is restricted to satisfy $E[s_1(x)] = 0$. Without this constraint, the intercept of the model would be unidentified.

The fit of the additive model is performed with the experimental procedure PROC GAM of SAS9 (SAS Institute Inc., Cary, NC, USA). This procedure adds a linear term to the non-parametrically modeled predictor that is fitted parametrically. By doing this, the effect of the predictor is divided into an overall linear trend and a non-parametric deviation of this linear trend. Thus, the applied model is

$$y = s_0 + \beta x + \tilde{s_1}(x) + \sum_{i=1}^{k} \beta_i Z_i + \varepsilon \ .$$

[1] In this work, generalized additive models are restricted to the special case of an identity link function between predictor an outcome. For this reason, the term *additive model* is used instead of *generalized additive models*.

Cubic smoothing splines are chosen for the function $\tilde{s_1}$. As described above, splines are piecewise polynomials that join at the edges (the knots) in a smooth way. A smoothing spline $f(x)$ is a spline function with a knot at each distinct data point, with the constraint of minimizing the term

$$\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \int f''(x)^2 dx \; . \tag{2.2}$$

The first part of (2.2) is the usual residual sum of squares (RSS), while the second part is a measure of "wiggliness" of the function. At first, it seems that the smoothing spline has too many parameters to be fit to the data, as it has as many knots as distinct data points. However, the restriction imposed by the smoothing parameter λ reduces the *effective number of parameters* of the smoothing spline. If $\lambda \to \infty$, the second part of (2.2) becomes dominating and $f(x)$ will reduce to a straight line because of its second derivative equal to zero. In this case the number of effective parameters would be one[1]. If $\lambda \to 0$, the RSS dominates (2.2) and $f(x)$ tends to interpolate the data points. Thus, the effective number of parameters would be that of a saturated model, i. e. equal to the number of observations.

The effective number of parameters of a smoothing spline is also referred to as its degrees of freedom (Hastie and Tibshirani, 1990). In general, this quantity is easier to interpret than λ, that depends on the unit of the predictor x. Therefore, the amount of flexibility of a smoothing spline is reported in general by its degrees of freedom.

A model selection procedure using the F-statistic (2.1) proposed in section 2.2.3 is given as a possibility for deciding about how much flexibility the smoothing spline should be given. The degrees of freedom for the smoothing spline is increased by 1 from one step to the next and the models are tested against each other. According to the model selection procedure for linear splines (see section 2.2), the selection procedure is stopped if the increase by one degree of freedom reveals no preference of the larger model in two consecutive model comparisons. The procedure is initialized by a normal linear regression on the interesting linear predictor. As explained above, this is equivalent to fitting an additive model with a smoothing spline of one degree of freedom.

The fit of the non-parametric function s_1 is performed by a back-fitting algorithm using the partial residuals $R_1 = y - s_0 - \beta x - \sum_{i=1}^{k} \beta_i Z_i$ as a basis for the estimates of s_1. Further details can be found in the SAS documentation of the GAM procedure (SAS, 2005) or in Breiman and Friedman (1985).

[1] In fact, a straight line has two parameters. But as the function is restricted in the additive model to a zero mean, the effective number of parameters reduces to one.

2.5 Analysis of Variance and Linear Regression

Two linear models are applied to compare the above presented methods with the standard methods, i. e. analysis of variance (ANOVA) and linear regression.

In order to perform ANOVA, the interesting continuous predictor is transformed into a class variable. The class boundaries are defined by the quartiles of the variable to create four groups of equal size. The number of groups are chosen to provide ANOVA with approximately the same amount of flexibility as for linear splines, fractional polynomials and additive models. The applied model is as follows:

$$y_i = \mu_0 + \alpha_i + \sum_{j=1}^{l} \beta_j C_j + \varepsilon \ ,$$

with y_i the dependent variable for the i^{th} group of the continuous predictor, μ_0 the overall mean, α_i the parameter for the i^{th} group and C_j the confounders with the corresponding parameters β_j.

For linear regression, the continuous predictor was directly included into the model. The resulting linear regression model is:

$$y = \mu_0 + \beta x + \sum_{j=1}^{l} \beta_j C_j + \varepsilon \ ,$$

with x the interesting continuous predictor and β its regression parameter.

2.6 Model Comparison

The goodness-of-fit of a model depends on the one hand, on its ability to explain the variability in the observed data, and on the other hand, on its simplicity which itself is often associated with easier interpretability. This becomes apparent by analysis of the perfect model fit of an interpolating saturated model that would clearly be too complex in most cases. In general, the aim is to identify a model that keeps a balance between data fit and simplicity. In order to achieve an appropriate measure of the goodness-of-fit, it is possible to penalize the fit to the data sample by the number of model parameters. Several goodness-of-fit measures have been proposed in the literature.

An easy way to describe the proportion of variance of the depending variable explained by the model is given by the coefficient of determination R^2. It is calculated by the residual sum

of squares (RSS) and the sum of squares of the data (SS):

$$R^2 = 1 - \frac{RSS}{SS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \ ,$$

with y_i the i^{th} observation of the depending variable, $\hat{y}_i$ the model prediction of the i^{th} observation and $\bar{y}$ the mean of the observations y_i. However, as this coefficient does not consider model complexity, other measures have to be considered as well.

The most commonly used measures of goodness-of-fit are presented below. A detailed comparison of these measures can be found in Burnham and Anderson (2004). The authors also make suggestions on how to decide which measure of goodness-of-fit should be chosen for model selection. However, as the emphasis in this work is rather on model comparison than on model selection, both presented information criteria will be calculated and compared.

2.6.1 Akaike Information Criterion

Akaike (1973) suggested a measure of goodness-of-fit based on information theory and maximum likelihood theory which is known as *Akaike Information Criterion* (AIC):

$$\text{AIC} = -2LL + 2p$$

where LL denotes the log-likelihood of the model and p the number of model parameters. The first part of the AIC is a measure of fit to the data, while the second can be interpreted as a penalizing term for the model complexity. In general, the smaller the AIC, the better the goodness-of-fit of the model.

In the late eighties, Hurvich and Tsai (1989) derived a small sample AIC (denoted AICC[1]) that has an additional penalizing term

$$\text{AICC} = -2LL + 2p + \frac{2p(p+1)}{n-p-1} \ ,$$

where n denotes the number of observations. As AICC $\rightarrow AIC$ if $n \rightarrow \infty$, Burnham and Anderson (2004) recommend that only the AICC be used unless $N/p > 40$. They also suggest using only AIC differences to the model with the minimum AIC

$$\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$$

with AIC_i the AIC of model i and AIC_{min} the minimum AIC of the studied models. "The Δ_i are easy to interpret and allow a quick 'strength of evidence' comparison and ranking of

[1]for AIC *corrected*

candidate hypotheses or models" (Burnham and Anderson, 2004). A rule of thumb is given for the interpretation of the models Δ_i. Models with

- $\Delta_i \leq 2$ show *substantial support (evidence)*,
- $4 \leq \Delta_i \leq 7$ show *considerably less support* and
- $\Delta_i > 10$ show *essentially no support*.

2.6.2 Bayesian Information Criterion

Another measure of goodness-of-fit was introduced by Schwarz (1978) and is generally referred to as the *Bayesian Information Criterion* (BIC) as it is derived from a Bayesian point of view:

$$BIC = -2LL + p\ln(n) \ ,$$

using the same notation as before. The name though may be misleading because BIC is not related to information theory (Burnham and Anderson, 2004).

BIC tends to chose models with less parameters in comparison to AIC, as for $n \geq 8$ the penalizing term becomes greater than in AIC.

The BIC has its origins in Bayes' theory. Therefore, if a-priori probabilities are specified for each studied model, a-posteriori probabilities can be derived via the $\Delta\mathrm{BIC}_i = \mathrm{BIC}_i - \mathrm{BIC}_{\min}$. Given the a-priori probabilities q_i of the i^{th} model, a-posteriori probabilities can be calculated as follows:

$$P(\mathrm{model}_i \,|\, \mathrm{data}) = \frac{\exp(-\frac{1}{2}\Delta\mathrm{BIC}_i)q_i}{\sum_j \exp(-\frac{1}{2}\Delta\mathrm{BIC}_j)q_j} \ .$$

For the calculations of a-posteriori probabilities in this work, vague priors are used, i. e. equal a-priori probabilities for each model in the set of considered models.

3 Results

3.1 Description of the Study Population

The complete dataset of the PAH study consisted of 368 male workers exposed to PAH in eight different industrial settings at 20 different companies in Germany. However, in some of the participating companies no measurements could be made at all and in others, only stationary PAH measurements could be made. These observations were therefore excluded from the analysis. The dataset with personal measurements of PAH exposure at the work place comprised 285 workers. In the following, this dataset is referred to as the study population.

None of the analyses presented in this section has been planned in advance. Therefore, all tests are of a purely explorative nature and have to be interpreted with care.

3.1.1 Description

The study was conducted in different industrial settings with occupational exposure to PAH. The number of workers by type of industry is shown in Table 3.1. Measurements were made in eleven German towns and 14 different companies. About 100 workers came from refractory and graphite electrode manufacturing industries, making up about two-thirds of the study population (35.1 % and 32.3 %). Sixty-three coke oven workers were included in the study population (22.1 %), while only 18 (6.3 %) and 12 (4.2 %) workers from tar distillation and converter infeed were available. Six of the 14 participating companies were located in North Rhine-Westphalia with a total of 109 workers (38.2 %).

The workers' age ranged from 19 to 62 years (data not shown). The mean age was 38.7 years with a standard deviation of 9.5 years. The median of 38 years was in good agreement with the mean, indicating a relatively symmetric age distribution. No information on age was available for 84 workers.

The nationality of the workers is also given in Table 3.1. Information on nationality was missing for 42 workers (14.7 %). Overall, 79.4 % of the workers with available information on nationality were German. Turkish was the second most frequent nationality with 34 workers (20.6 %). The rest of the workers came from different European countries except for one Moroccan worker.

Table 3.1: Number of Workers by Type of Industry, Nationality and Smoking Status (N = 285)

Variable	**N**	(%)
Type of industry		
Missing	0	
Coke oven	63	(22.1)
Converter infeed	12	(4.2)
Manufacture of graphite electrodes	92	(32.3)
Manufacture of refractory	100	(35.1)
Tar distillation	18	(6.3)
Nationality		
Missing	42	
German	193	(79.4)
Other	50	(20.6)
Smoking status		
Missing	41	
Never-smoker	56	(23.0)
Former smoker	32	(13.1)
Current smoker	156	(63.9)
Smoking status (derived)		
Missing	1	
Current non-smoker	100	(35.2)
Current smoker	184	(64.8)

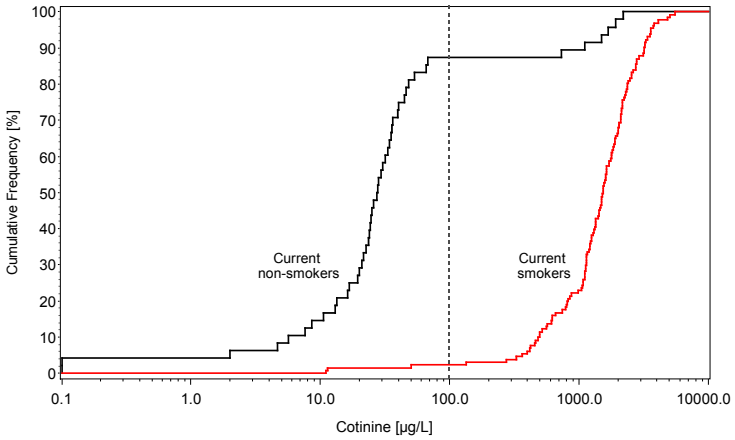
Smoking habits of the workers were assessed in two ways. Firstly, the workers were asked in the questionnaire to classify themselves as former, current or non-smokers. This data is shown in Table 3.1. The vast majority (156 workers, 63.9 %) answered that they currently smoked. Thirty-two workers had smoked in the past (13.1 %), and 56 reported that they had never smoked (23.0 %). Information on the smoking status was missing from the questionnaire for 41 workers. The amount of smoking was not assessed by the questionnaire.

Secondly, the workers cotinine concentration was measured in urinary samples. However, this information was not available for 66 workers. The cumulative distribution function of cotinine for self-assessed current non-smokers[1] and current smokers is shown in Figure 3.1.

In order to provide consistent information of smoking status for the whole study population, a combination of cotinine concentration and information from the questionnaire was used to classify the workers into current smokers and current non-smokers. Classification using a cut-off of 100 μg/L on cotinine concentration revealed a good concordance with the questionnaire (93.3 % for current non-smokers and 95.5 % for current smokers). Where both pieces of information were available, classification using the cut-off on cotinine was preferred

[1] Never-smokers and former smokers were grouped together as current non-smokers.

Figure 3.1: Empirical Cumulative Distribution of Cotinine Levels [μg/L] for Self-Assessed Current Non-Smokers (Never-Smokers and Former Smokers) and Current Smokers and Applied Cut-Off for Classification of Smoking Status 100 μg/L



to the questionnaire. Cotinine concentration was used to derive the smoking status in 219 cases (77.1 %), and information from the questionnaire was used in 65 cases (22.9 %). The derived smoking status is also shown in Table 3.1. One hundred workers were classified as current non-smokers (35.2 %) and 184 as current smokers (64.8 %). One worker could not be classified due to missing data. As the current smoking status was considered as confounder and therefore included in all models, only 284 workers with complete data could be analyzed.

3.1.2 Airborne Exposure and Urinary Metabolites

As mentioned in section 2.1.3, air sampling was performed in the worker's breathing zone for an average of two hours. The concentration of sixteen US EPA PAHs was determined in the air samples. Table 3.2 describes the distribution of the sum of these sixteen PAHs, phenathrene (PHE) and some selected PAHs. In the same manner, the table shows the distribution of urinary metabolites of PHE and pyrene. The distribution of all these variables were highly skewed and seemed to be log-normal. Consequently, geometric means and geometric standard deviations are presented.

The carcinogenic compounds such as benzo[a]pyrene had a relevant fraction of measurements below LQ. PHE was a prominent PAH compound with only five measurements below LQ. For all variables, geometric means corresponded well to the median values supporting the chosen log-transformation for further analyses.

In order to examine the associations between PHE and the remaining PAHs, Spearman rank correlation coefficients were calculated (Table 3.3). Overall, PHE correlated strongly with the other compounds and their sum ($r_S = 0.80$, $P < 0.0001$). The correlation with

Table 3.2: Distribution of the Study Variables and Selected Exposure Variables ($N = 285$)

				Percentiles				
Variable	$\mathbf{N_{\leq LQ}}$[a]	**GM**[b]	**GSD**[c]	**5 %**	**25 %**	**50 %**	**75 %**	**95 %**
Sum of 16 EPA PAHs [$\mu g/m^3$]	–	34.8	4.44	3.23	14.2	30.2	90.2	531
Dibenz(a,h)anthracene [$\mu g/m^3$]	152	0.07	7.06	0.01	0.01	0.08	0.28	1.70
Benzo[a]pyrene [$\mu g/m^3$]	68	0.38	7.03	0.02	0.09	0.41	1.42	11.9
Phenanthrene [$\mu g/m^3$]	5	4.82	4.87	0.34	1.79	5.15	13.1	65.4
Sum of OH-phenanthrenes [$\mu g/g\,crn$]	0	9.53	3.19	1.51	4.25	10.1	21.2	64.3
1-OH-pyrene [$\mu g/g\,crn$]	0	5.25	3.90	0.51	2.17	5.62	12.6	38.5

[a]Number of observations below the limit of quantification (LQ) (set to half of the LQ); [b]Geometric mean; [c]Geometric standard deviation

acenaphthylene is not presented, because of the high number of measurements below the LQ. The low correlation coefficients of PHE with several substances such as benzo[a]pyrene and dibenz[a,h]anthracene may also partly be due to the large number of measurements below the LQ for those compounds. These correlations became stronger when the calculation was restricted to data with measurements above LQ.

Table 3.3: Spearman Rank Correlations of Phenanthrene Exposure in Air during a Working Shift with other US EPA PAHs ($N = 285$)

PAH	$\mathbf{r_S}$[a]	$\mathbf{N_{>LQ}}$[b]	$\mathbf{r_{S,>LQ}}$[c]	**PAH**	$\mathbf{r_S}$	$\mathbf{N_{>LQ}}$	$\mathbf{r_{S,>LQ}}$
Sum of 15 PAH	0.80	–	–	Benz[a]anthracene	0.42	229	0.41
Anthracene	0.95	269	0.94	Benzo[k]fluoranthene	0.34	197	0.39
Fluoranthene	0.81	252	0.74	Benzo[b]fluoranthene	0.34	211	0.36
Fluorene	0.79	220	0.81	Benzo[a]pyrene	0.34	217	0.34
Pyrene	0.70	237	0.60	Indeno[1,2,3-cd]pyrene	0.27	185	0.36
Acenaphthene	0.54	125	0.77	Dibenz[a,h]anthracene	0.24	133	0.60
Naphthalene	0.49	209	0.44	Benzo[ghi]perylene	0.24	147	0.53
Chrysene	0.45	230	0.45	Acenaphthylene[d]	–	45	–

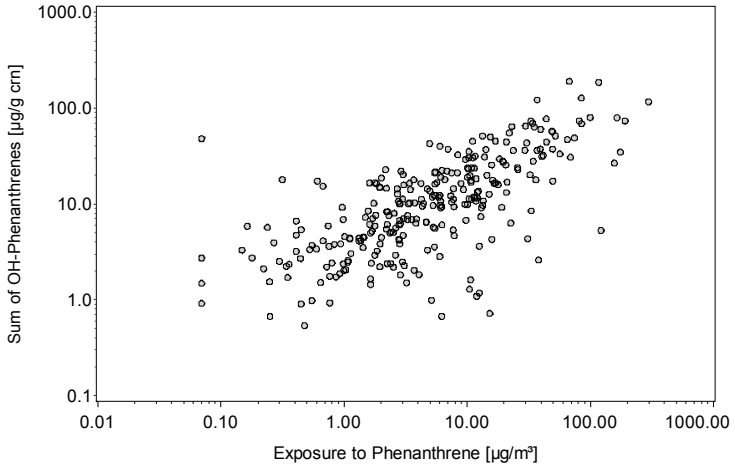
[a]Spearman rank correlation coefficients (all correlations with $P < 0.0001$); [b]Number of observations above the limit of quantification (LQ); [c]Spearman rank correlation coefficients for values above the LQ (all correlations with $P < 0.0001$); [d]Correlations not calculated because 84.2 % of measurements were below LQ

3.2 Dose-Response Analysis

In order to describe the dose-response relation between phenanthrene exposure in the workplace (PHE) and urinary sum of 1-, 2-+9-, 3- and 4-OH-phenanthrene (OHPHE), the methods described in detail in the sections 2.2, 2.3 and 2.4 were applied. The results of these analyses are presented below. Additionally, ANOVA and linear regression were carried out to compare the results of linear splines, fractional polynomials and additive models with standard methods. Due to the highly skewed distribution of OHPHE, the variable was log-transformed. Parameter estimates are back-transformed to the original scale by the exponential function.

Figure 3.2 shows the scatterplot of exposure to PHE and OHPHE with both variables on the logarithmic scale. Examination of the plot as well as Spearman's rank correlation reveals a strong monotonic association between the variables ($r_S = 0.70$, $P < 0.0001$).

Figure 3.2: Scatterplot of Exposure to Phenanthrene in the Workplace and Sum of OH-Phenanthrenes in Urine on Log-Scales ($N = 285$)



Type of industry and current smoking were included as confounders in all models. For reasons of completeness, parameter estimates for the confounders are presented in section 3.2.1. However, as these results are not of primary interest and do not change largely between the applied models, they are not presented in the following sections.

3.2.1 Analysis of Variance

As described in section 2.5, the exposure variable PHE was transformed into a class variable by using quartiles as class boundaries (see Table 3.2). Four groups of equal size are thus established.

The resuls of the ANOVA are given in Table 3.4. The intercept, corresponding to the estimated level for the lowest exposure group, was 3.85 μg/g crn (95 % CI 2.92–5.08 μg/g crn).

A test of the intercept against a given value is meaningless and therefore was not performed. The overall F-test for an effect of the grouped exposure variable was highly significant ($P < 0.0001$). Comparisons of the higher levels of exposure against the lowest level (PHE<1st quartile) revealed clear differences in the amount of excretion of OHPHE. Workers with an exposure between the 1st and 2nd quartile of PHE (group 2) and between the 2nd and 3rd quartile (group 3) showed a 2.13 and 2.79 times higher excretion of OHPHE than workers in the lowest exposure group (95 % CI 1.61–2.81 and 2.10–3.72, respectively; both $P < 0.0001$). Group 4 (PHE>3rd quartile) exhibited 7.84 times higher values than group 1 (95 % CI 5.83–10.5, $P < 0.0001$).

Table 3.4: Results of Analysis of Variance with Grouped Exposure to PHE (4 Groups Defined by Quartiles)

Variable	**DF**[a]	**exp(β)**[b]	**95 % CI**[c]	***P***
Intercept	1	3.85	(2.92, 5.08)	–
Exposure group[d]	3	–	–	<0.0001
Group 2 vs. group 1	1	2.13	(1.61, 2.81)	<0.0001
Group 3 vs. group 1	1	2.79	(2.10, 3.72)	<0.0001
Group 4 vs. group 1	1	7.84	(5.83, 10.5)	<0.0001
Type of industry[e]	4	–	–	0.001
CO vs. GE[f]	1	1.53	(1.16, 2.02)	0.003
CV vs. GE	1	2.01	(1.19, 3.41)	0.01
RE vs. GE	1	1.59	(1.24, 2.05)	0.0003
TD vs. GE	1	1.05	(0.69, 1.61)	0.82
Current Smoking	1	0.97	(0.79, 1.19)	0.75

[a]Degrees of freedom; [b]Back-transformed parameter estimate; [c]Confidence interval of exp(β); [d]Group 1: PHE<1st quartile, Group 2: 1st quartile≤PHE<2nd quartile, Group 3: 2nd quartile≤PHE<3rd quartile, Group 4: 3rd quartile≤PHE; [e]CO=Coke oven, CV=Converter, GE=Graphite electrodes, RE=Refractory, TD=Tar distillation; [f]Lowest group (GE) chosen as reference

Examination of the overall F-tests of the included confounder variables revealed that the type of industry influenced the amount of excretion of OHPHE ($P = 0.001$). Comparing the different industrial settings against production of graphite electrodes showed higher values of OHPHE for coke oven workers (exp(β) = 1.53, 95 % CI 1.16–2.02, $P = 0.003$), converter infeed workers (exp(β) = 2.01, 95 % CI 1.19–3.41, $P = 0.01$) and refractory production workers (exp(β) = 1.59, 95 % CI 1.24–2.05, $P = 0.0003$). Values of OHPHE for tar distillation were at the same level as for graphite electrodes (exp(β) = 1.05, 95 % CI 0.69–1.61, $P = 0.82$). Current smoking had no significant impact an the excretion of OHPHE in urine (exp(β) = 0.97, 95 % CI 0.79–1.19, $P = 0.75$).

Figure 3.3 shows the modeled association between the grouped exposure to PHE and the urinary excretion of OHPHE both on the logarithmic scale, together with the confounder

adjusted observations. The standardized residuals of the model and the plot of the quantiles of the residuals versus the quantiles of the standard normal distribution (QQ plot) are shown in Figure 3.4. The lower and upper dashed line in the plot of standardized residuals indicate the 2.5 % and 97.5 % quantile of the standard normal distribution. Seventeen residuals were observed outside of this range (6.0 %). Examination of the QQ plot reveals the assumption of a normal distribution of the residuals to be quite reasonable. Slight deviations from normality have to be noticed in the left tail of the distribution only.

Figure 3.3: Best Fit Model using Analysis of Variance with Grouped Exposure (4 Groups Defined by Quartiles), 95 % Confidence Band and Confounder Adjusted Data

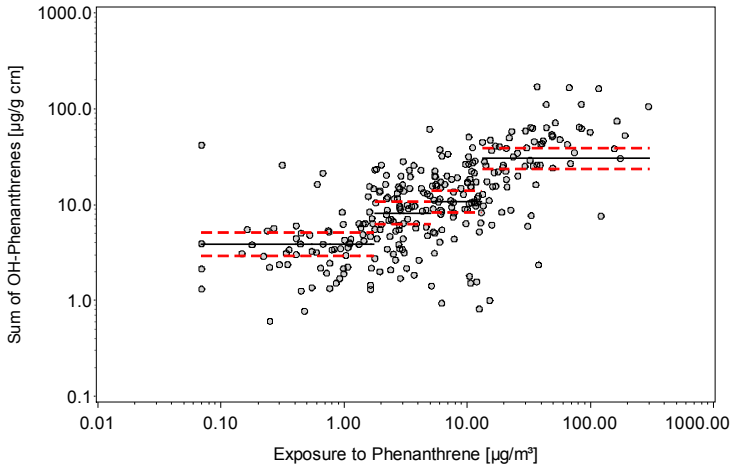
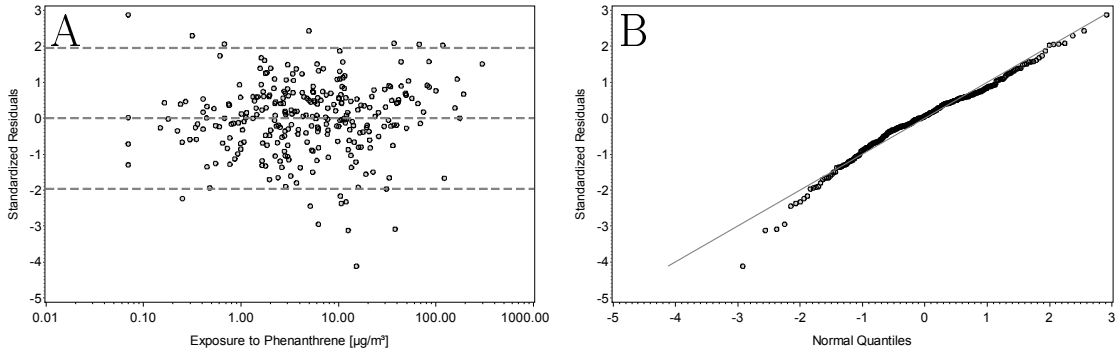


Figure 3.4: Standardized Residuals (A) and QQ Plot (B) of the Best Fit Model using Analysis of Variance with Grouped Exposure (4 Groups Defined by Quartiles)



3.2.2 Linear Regression

For linear regression, the exposure to PHE was itself log-transformed and included in the model as continuous predictor. The logarithm to the basis 2 was chosen for this transformation making it easier to interpret the regression parameter. Transformation of the corresponding regression parameter by the exponential function allows it to be interpreted as the change of OHPHE under a doubling of PHE (see Appendix B). The parameter estimates are given in Table 3.5. The level of OHPHE under an exposure to PHE of $1\,\mu g/m^3$ is given by the intercept and was $4.98\,\mu g/g\,crn$ with a 95 % CI of 3.95–$6.29\,\mu g/g\,crn$. As for ANOVA, a test for this parameter is meaningless and was not performed. Under a doubling of the exposure to PHE, the excretion of OHPHE increased by a factor of 1.38 (95 % CI 1.32–1.45, $P < 0.0001$).

Table 3.5: Results of Linear Regression with Log-Transformed Exposure to PHE

Variable	**DF**[a]	**exp(β)**[b]	**95 % CI**[c]	***P***
Intercept	1	4.98	(3.95, 6.29)	–
$\log_2$(PHE)	1	1.38	(1.32, 1.45)	<0.0001

[a]Degrees of freedom; [b]Back-transformed parameter estimate adjusted for type of industry and current smoking; [c]95 % confidence interval of exp(β)

The corresponding model, 95 % confidence bands and confounder adjusted data are shown in Figure 3.5. The resulting standardized residuals and the QQ plot for an analysis of the normality are given in Figure 3.6. As already seen with ANOVA, the left tail of the distribution shows some deviations from normality. However, the assumption of normality seems reasonable.

Figure 3.5: Best Fit Model using Linear Regression with Log-Transformed Exposure as Predictor, 95 % Confidence Band and Confounder Adjusted Data

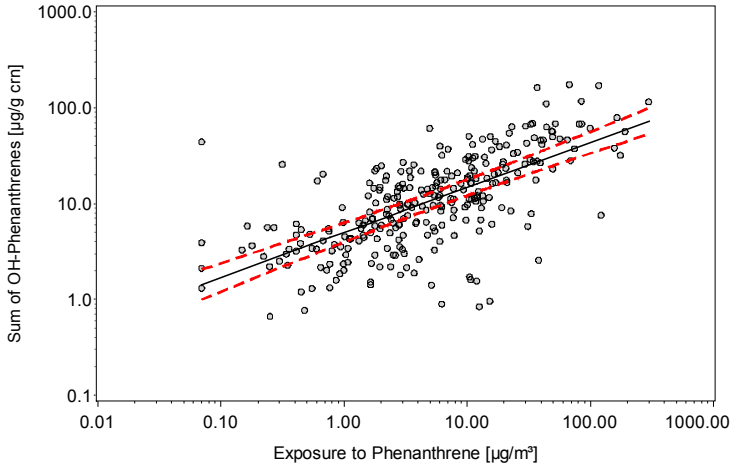
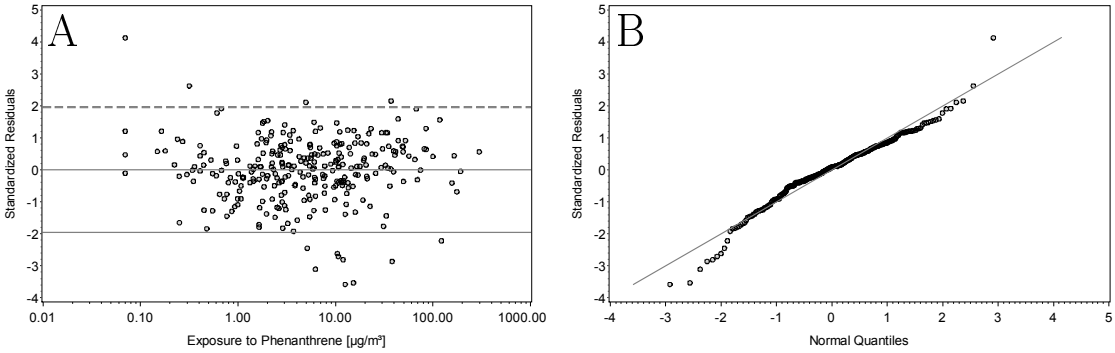


Figure 3.6: Standardized Residuals (A) and QQ Plot (B) of the Best Fit Model using Linear Regression with Log-Transformed Exposure as Predictor



3.2.3 Linear Splines

The linear splines model presented in this section was achieved using the DTL basis for the space of linear splines given in section 2.2.1.3. Therefore, estimates of the slope in each segment are yield by the parameter estimates corresponding to the DTL function with a positive slope in that segment.

The procedure for identifying the number of knots of the linear splines model is shown in Table 3.6. As described in section 2.2.3, the number of knots was increased by one until the inclusion of two subsequent knots did not result in a better model fit. For each number of knots, the best location of the knots was determined. A grid of 200 points was used to initiate the search to avoid local maxima at each stage. In order to avoid overfitting, no models with fewer than 10 % of data points, i. e. at least 29 observations, between two subsequent knots were allowed.

Table 3.6: Model Selection Procedure for Linear Splines

			Test[a]		
Model	**RSS**[b]	**DF$_{\text{Model}}$**[c]	**F**	**DF**[c]	***P***
No knots	194.1	277			
One knot	184.3	275	7.33	2	0.001
Two knots	183.3	273	0.73	2	0.48
Three knots	181.0	271	1.76	2	0.17

[a]F-test versus previous model; [b]Residual sum of squares; [c]Degrees of freedom

The residual sum of square (RSS) for the model without knots[1] was 194.1 with 277 degrees of freedom (DF). By including one knot into the model, the RSS decreased to 184.3 with

[1]In fact, this corresponds to the model described in section 3.2.2.

275 DF. Comparing both models, the F-test revealed a significantly better fit with one knot versus no knots ($P = 0.001$).

The inclusion of a second knot resulted in a RSS of 183.3 with 273 DF[1]. The F-test against the model with one knot showed no advantage of the inclusion of the second knot into the model ($P = 0.48$). Albeit this result, the model fit procedure was pursued.

Inclusion of a third knot yielded a RSS of 181.0 with 271 DF. The model was compared against the previous model with two knots. Again, the F-test showed no advantage of the model with three knots ($P = 0.17$). Hence, the model fit procedure was stopped and the model with one knot was chosen.

Here, the F-test was applied in order to select the appropriate model. This procedure can be verified by means of the information criteria described in section 2.6. Results of the information criteria and derived quantities are shown in Table 3.7. The variance explained by the model is given by the coefficient of determination (R^2). It increased from 49.1 % with linear regression (no knots) to 52.6 % with the three knot model. As the models are nested, R^2 is an increasing function by the number of knots. For model comparison, the corrected Akaike and the Bayesian information criterion (AICC, BIC) were calculated. For both measures, the one knot model revealed the smallest values and hence the greatest support. Using the rule of thumb for the interpretation of the differences in AICC, the model without knots showed a difference of 10.4 and hence *essentially no support*. The two and three knot model showed differences of 2.9 and 3.7, i. e. *less support*. The comparison of the models based on the BIC yield slightly slightly different results, because of the BIC's stronger preference of small models. The two and three knot model both had a-posteriori probabilities below 1 % (0.6 % and 0.0 %) indicating only little support. Linear regression had an a-posteriori probability of 15.3 %, while the one knot model showed the largest probability with 84.0 %.

Table 3.7: Model Fit and Information Criteria for the Applied Linear Spline Models

Model	p[a]	$\mathbf{R^2}$[b]	**AICC**[c]	Δ_i[d]	**BIC**[e]	Δ**BIC**$_i$[f]	$P_{\mathbf{post}}$[g]
No knots	1	49.1 %	712.4	10.4	737.5	3.4	15.3 %
One knot	3	51.7 %	702.0	0.0	734.1	0.0	84.0 %
Two knots	5	52.0 %	704.8	2.9	744.0	9.9	0.6 %
Three knots	7	52.6 %	705.6	3.7	751.7	17.6	0.01 %

[a]Number of model parameters used for the exposure variable; [b]Coefficient of determination; [c]Akaike information criterion (corrected); [d]$\Delta_i = \text{AICC}_i - \text{AICC}_{\text{min}}$; [e]Bayesian information criterion; [f]$\Delta\text{BIC}_i = \text{BIC}_i - \text{BIC}_{\text{min}}$; [g]a-posteriori probability

Results of the model fit using linear splines with one knot are given in Table 3.8. The

[1]At each inclusion of an additional knot, the DF are reduced by two. One DF is used for the additional knot, the other for the slope in the additional segment.

level of OHPHE under an exposure to PHE of 1 $\mu g/m^3$ (intercept) was 3.45 $\mu g/g\,crn$ (95 % CI 2.00–5.97 $\mu g/g\,crn$). The identified knot location (back-transformed on the original scale of PHE) was 0.77 $\mu g/m^3$ with a 95 % CI of 0.35–1.68 $\mu g/m^3$. In the first segment of the exposure to PHE up to the concentration of 0.77 $\mu g/m^3$ identified by the knot, no significant influence of PHE on OHPHE was found ($P = 0.38$). OHPHE decreased by an estimated factor of 0.96 under a doubling of PHE (95 % CI 0.73–1.26).

In the second segment above the identified knot, a clear influence of PHE on the excretion of OHPHE was observed. Under a doubling of PHE, OHPHE increased by a factor of 1.47 (95 % CI 1.39–1.56, $P < 0.0001$).

Table 3.8: Results of Linear Splines with Log-Transformed Exposure to PHE

Variable	**DF**[a]	**exp(β)**[b]	**95 % CI**[c]	***P***
Intercept	1	3.45	(2.00, 5.97)	–
Knot [$\mu g/m^3$]	1	0.77	(0.35, 1.68)	–
$\log_2$(PHE)<$\log_2$(knot)	1	0.96	(0.73, 1.26)	0.38
$\log_2$(PHE)>$\log_2$(knot)	1	1.47	(1.39, 1.56)	<0.0001

[a]Degrees of freedom; [b]Back-transformed parameter estimate adjusted for type of industry and current smoking; [c]Confidence interval of exp(β)

Figure 3.7 shows the resulting model, 95 % confidence bands and confounder adjusted data. The discontinuity of the confidence bands is a consequence of the model's discontinuity and the non-existence of derivatives at the knot location. Figure 3.8 shows the resulting residuals as well as the QQ plot for the normality check. As already seen for ANOVA and linear regression, the plot indicates a heavy left tail of the distribution. However, with exception of this feature the assumption of normality seems reasonable.

Figure 3.7: Best Fit Model using Linear Splines with Log-Transformed Exposure as Predictor, 95 % Confidence Band and Confounder Adjusted Data

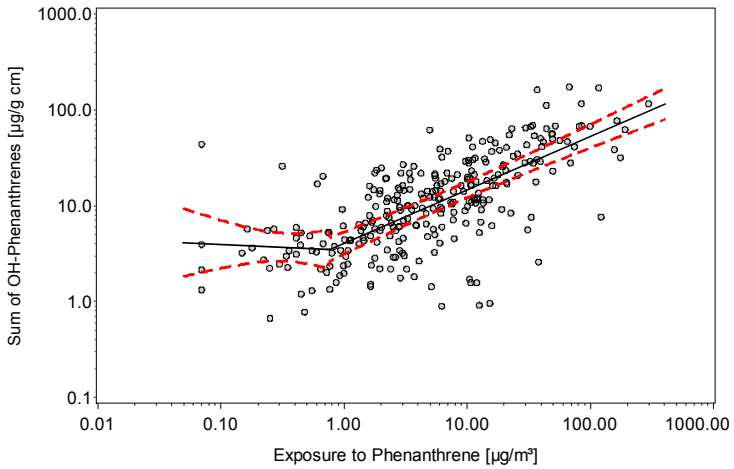
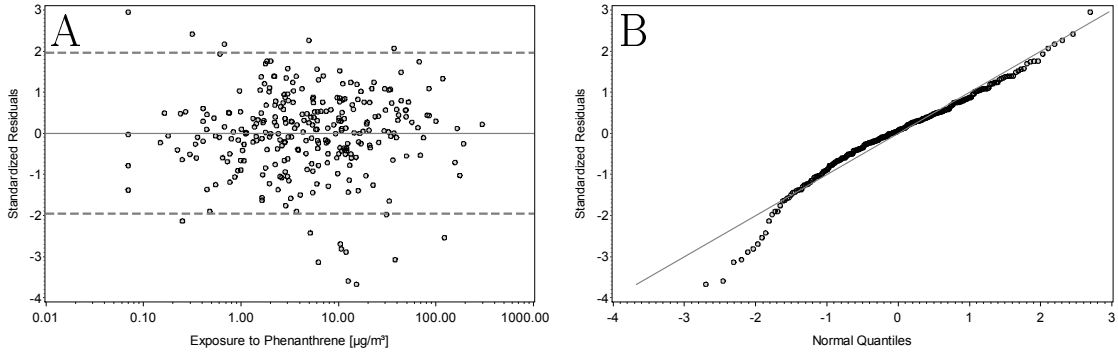


Figure 3.8: Standardized Residuals (A) and QQ Plot (B) of the Best Fit Model using Linear Splines with Log-Transformed Exposure as Predictor



3.2.4 Fractional Polynomials

When applying fractional polynomials, reflection is needed about how to deal with the exposure variable in the models. For the other methods described in this chapter, exposure to PHE was log-transformed. The same approach may be chosen to enable an easy comparison between the other methods and fractional polynomials. However, as log-transformation is already comprised in the considered transformations by fractional polynomials, this approach could be questioned. Therefore, fractional polynomials will be applied twice: once with the original untransformed exposure to PHE as continuous predictor and twice with the log-transformed exposure variable. The former approach is shown in section 3.2.4.1, the latter in section 3.2.4.2.

3.2.4.1 Untransformed Exposure

The usual set of transformations was modified slightly to be consistent with the other methods. The natural logarithm in case of $p = 0$ was replaced by the logarithm to the basis 2. The same replacement was applied for second order fractional polynomials in case of $p_1 = p_2$. The usual second transformation $\ln(x)x^{p_1}$ was replaced by $\log_2(x)x^{p_1}$.

Table 3.9 gives the results of the model selection procedure of fractional polynomials. All possible deviance differences for first and second order fractional polynomials are calculated. Deviance differences for the first order fractional polynomials are calculated against the linear model, i. e. $p = 1$. The largest deviance achieved for each order is underlined.

The best first order model was accomplished with $p = 0$, i. e. the logarithmic transformation, with a deviance difference against the linear model of 81.8. The χ^2-test of the model with logarithmic transformation of PHE versus the linear model showed that the hypothesis $p = 1$ can be rejected ($P < 0.0001$).

Table 3.9: Model Selection Procedure for Fractional Polynomials with Untransformed Exposure to PHE

Model	p_2 \ p_1	Deviance differences							
		−2	−1	−0.5	0	0.5	1	2	3
First degree	–	−63.8	−47.5	−0.9	81.8	63.2	0.0	−47.2	−57.3
Second degree	−2	−119.3	−85.3	−39.4	11.9	−18.5	−79.2	−125.3	−135.3
	−1		−52.8	−16.5	14.5	−15.8	−67.7	−109.7	−119.2
	−0.5			4.1	14.4	−7.9	−39.6	−67.81	−74.6
	0				10.9	5.8	2.3	0.4	0.2
	0.5					12.9	11.4	1.4	−6.1
	1						−5.4	−39.3	−56.7
	2							−90.7	−108.5
	3								−123.3

The best second order model with a deviance difference of 14.5 was identified for $p_1 = -1$ and $p_2 = 0$, i. e. PHE^{-1} and $\log_2$(PHE). Deviance differences for the second order fractional polynomials were calculated against the best first order model, i. e. the $\log_2$-transformation $p = 0$. The χ^2-test with two DF revealed that the second order model should be preferred ($P = 0.001$).

The model selection procedure is inspected by means of the information criteria of section 2.6 (see Table 3.10). The model presented in the first line of the table is actually equivalent to the linear regression model presented in section 3.2.2, the linear splines model without knots and the additive model with smoothing spline of 1 degree of freedom (see Table 3.7 and 3.16). The only difference is in the number of parameters. The reason for this is that for the first degree model, the linear regression is the result of the model selection procedure and thus the linear term has an exponent equal to 1 as additional parameter. Consequently, the values for AICC and BIC differ because of their inclusion of the number of parameters for calculation.

Table 3.10: Model Fit and Information Criteria for the Applied Fractional Polynomial Models with Untransformed Exposure

Model	$\boldsymbol{p}$[a]	$\mathbf{R^2}$[b]	**AICC**[c]	Δ_i[d]	**BIC**[e]	Δ**BIC**$_i$[f]	$\boldsymbol{P}_{\mathbf{post}}$[g]
First order	2	49.1 %	714.6	10.1	743.2	3.1	17.6 %
Second order	4	51.7 %	704.4	0.0	740.1	0.0	82.4 %

[a]Number of model parameters used for the exposure variable; [b]Coefficient of determination; [c]Akaike information criterion (corrected); [d]Δ_i = AICC$_i$ − AICC$_{\min}$; [e]Bayesian information criterion; [f]ΔBIC$_i$ = BIC$_i$ − BIC$_{\min}$; [g]a-posteriori probability

The variance explained by the model increased from first order model to the second order by 2.6 percentage points. On the basis of both information criteria AICC and BIC, the second

order model revealed the better model fit and greatest support. The first order model showed *essentially no support*, with a difference in AICC of 10.1. Similarly, the second order model is suggested by a-posteriori probabilities derived from differences in BIC, though they did not reveal such a large preference (17.6 % against 82.4 %).

Parameter estimates of the second order model are shown in Table 3.11. The level of OHPHE under an exposure to PHE of $1\,\mu g/m^3$ (intercept) was $3.87\,\mu g/g\,crn$ with a 95 % CI of 2.98–5.03 $\mu g/g\,crn$. The inverse of PHE showed a highly significant impact on the excretion of OHPHE ($P = 0.0002$). The transformed estimate of the regression parameter of PHE^{-1} was 1.14 with a 95 % CI of 1.06–1.21. The second transformation $\log_2(PHE)$ also showed a clear influence on OHPHE ($P < 0.0001$). The transformed estimate of the regression parameter was 1.48 (95 % CI 1.40–1.57).

Table 3.11: Results of Fractional Polynomials with Untransformed Exposure to PHE

Variable	**DF**[a]	**exp(β)**[b]	**95 % CI**[c]	***P***
Intercept	1	3.87	(2.98, 5.03)	–
PHE^{-1}	1	1.14	(1.06, 1.21)	0.0002
$\log_2(PHE)$	1	1.48	(1.40, 1.57)	<0.0001

[a]Degrees of freedom; [b]Back-transformed parameter estimate adjusted for type of industry and current smoking; [c]Confidence interval of exp(β)

However, the interpretation of these results is not straightforward. As PHE^{-1} is a decreasing function in PHE, an estimate >1 indicates a decrease of OHPHE with increasing PHE. The transformed parameter of $\log_2$ can be interpreted as before as the factor of alteration under a doubling of PHE. Nonetheless, the estimates cannot be interpreted individually, but only together. Therefore, it has to be noticed that the influence of the term PHE^{-1} becomes smaller if PHE increases. Thus for large values of PHE the influence of $\log_2(PHE)$ becomes dominating and the corresponding transformed regression parameter can be interpreted as described in Appendix B. Consequently, it can be stated that for larger values of PHE, OHPHE increases approximately by a factor of 1.48 under a doubling of PHE.

A direct way of interpreting is given by an illustration of the dose-response curve which is shown in Figure 3.9 together with the 95 % confidence bands and the confounder adjusted data points. It can be seen that the influence of PHE^{-1} is mainly present for smaller values of PHE while for larger values the model becomes linear on a logarithmic scale due to the influence of $\log_2(PHE)$. Figure 3.10 shows the resulting standardized residuals and their QQ plot. The left tail of the distribution shows some deviations from the diagonal indicating normality. Nonetheless, the assumption of normality seems reasonable.

Figure 3.9: Best Fit Model using Fractional Polynomials with Untransformed Exposure as Predictor, 95 % Confidence Band and Confounder Adjusted Data

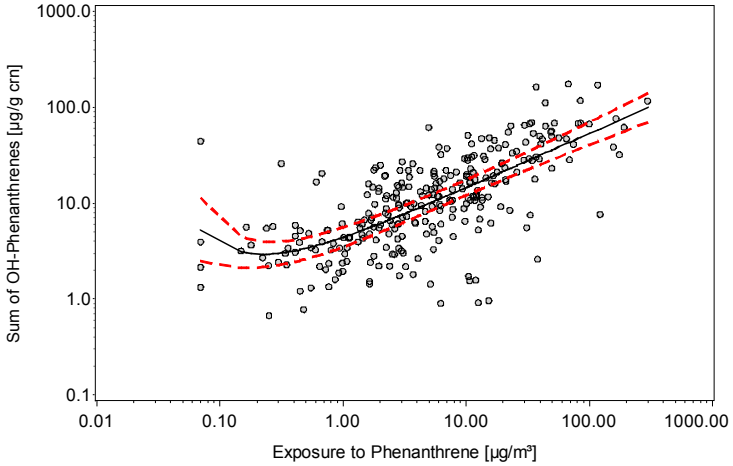
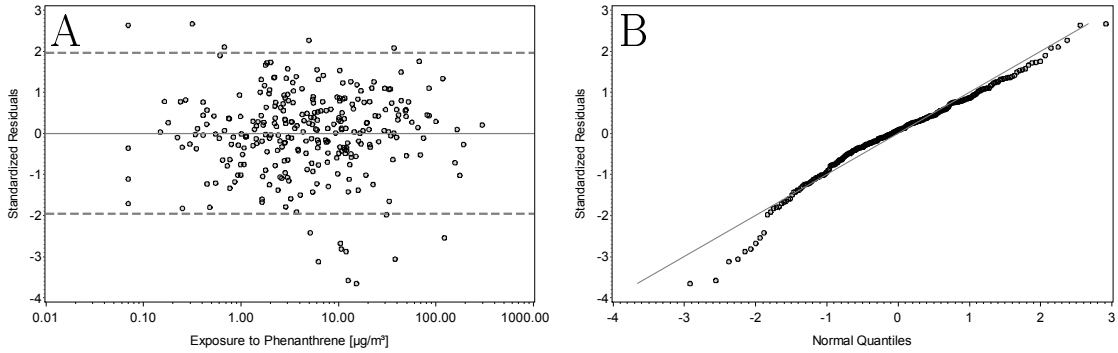


Figure 3.10: Standardized Residuals (A) and QQ Plot (B) of the Best Fit Model using Fractional Polynomials with Untransformed Exposure as Predictor



3.2.4.2 Log-Transformed Exposure

As many values of PHE are lower than $1\,\mu g/m^3$, application of the $\log_2$ results in non-positive values for the predictor variable. In order to avoid this problem, PHE is multiplied by the factor 1000 and thus is expressed in terms of the unit ng/m^3. This is equivalent to adding the constant $\log_2(1000)$ to $\log_2(\text{PHE})$.

The results of the model selection procedure are shown in Table 3.12. The deviance differences for the first order fractional polynomials are calculated against the linear model, i. e. $p = 1$. The largest deviance achieved for each order is underlined.

The best model fit for a first order fractional polynomial was achieved with $p = 3$, i. e. the cubic transformation, with a deviance difference against the linear model of 9.3. The χ^2-test of the model with cubic transformation of PHE versus the linear model showed that the hypothesis $p = 1$ can be rejected ($P = 0.003$).

Table 3.12: Model Fit Procedure for Fractional Polynomials with Log-Transformed Exposure to PHE

Model	p_2 \ p_1	Deviance differences							
		-2	-1	-0.5	0	0.5	1	2	3
First degree	–	−71.9	−44.2	−30.8	−18.6	−8.2	0.0	9.1	<u>9.3</u>
Second degree	−2	−0.8	3.0	4.3	5.1	5.48	5.3	3.5	0.0
	−1		4.9	5.3	<u>5.51</u>	5.3	4.8	2.9	0.0
	−0.5			5.50	5.4	5.0	4.4	2.6	0.1
	0				5.1	4.6	4.0	2.3	0.2
	0.5					4.1	3.5	2.0	0.3
	1						2.9	1.7	0.4
	2							1.2	0.8
	3								1.1

The best second order model with a deviance difference of 5.5 was identified for $p_1 = -1$ and $p_2 = 0$, i. e. $\big(\log_2(\text{PHE})\big)^{-1}$ and $\ln\big(\log_2(\text{PHE})\big)$. Deviance differences for the second order fractional polynomials were calculated against the best first order model, i. e. the cubic transformation $p = 3$. The χ^2-test with two DF revealed that the first order model should be preferred ($P = 0.07$).

Again, the model selection procedure is checked by the calculation of information criteria and the derived quantities. These results are presented in Table 3.13. The variance explained by the models increased from first order model to second order model from 50.8 % to 51.7 %. In contrast to the results of fractional polynomials with the untransformed exposure variable, different models were suggested by AICC and BIC. The best model fit indicated by AICC was the second order model, whereas the first order model had the better model fit according to BIC. However, despite the second order model having the better model fit by AICC, the rule of thumb indicates a *substantial support* of the first order model. The a-posteriori probabilities calculated from differences in BIC, clearly preferred the first order model with a probability of 94.9 % against 5.1 % for the second order.

Table 3.13: Model Fit and Information Criteria for the Applied Fractional Polynomial Models with Log-Transformed Exposure to PHE

Model	$\boldsymbol{p}$[a]	$\mathbf{R^2}$[b]	**AICC**[c]	Δ_i[d]	**BIC**[e]	$\Delta\mathbf{BIC}_i$[f]	$\boldsymbol{P}_{\mathbf{post}}$[g]
First order	2	50.8 %	705.3	1.2	734.0	0.0	94.9 %
Second order	4	51.7 %	704.1	0.0	739.8	5.9	5.1 %

[a]Number of model parameters used for the exposure variable; [b]Coefficient of determination; [c]Akaike information criterion (corrected); [d]$\Delta_i = \text{AICC}_i - \text{AICC}_{\text{min}}$; [e]Bayesian information criterion; [f]$\Delta\text{BIC}_i = \text{BIC}_i - \text{BIC}_{\text{min}}$; [g]a-posteriori probability

Table 3.14: Results of Fractional Polynomials with Log-Transformed Exposure to PHE

Variable	**DF**[a]	**exp(β)**[b]	**95 % CI**[c]	***P***
Intercept	1	5.00	(3.99, 6.28)	<0.0001
$\log_2(\text{PHE})^3$	1	1.0007	(1.0006, 1.0008)	<0.0001

[a]Degrees of freedom; [b]Back-transformed parameter estimate adjusted for type of industry and current smoking; [c]Confidence interval of exp(β)

Parameter estimates of the first order model are shown in Table 3.14. The level of OHPHE under an exposure to PHE of 1 μg/m^3 (intercept) was 5.00 μg/g crn with a 95 % CI of 3.99–6.28 μg/g crn. The cubic transformation of $\log_2$(PHE) showed a highly significant impact on the excretion of OHPHE ($P < 0.0001$). The transformed estimate of the regression parameter was 1.0007 with a 95 % CI of 1.0006–1.0008.

As seen before for the untransformed exposure variable, the easiest way of interpreting of these results is given by an illustration of the dose-response curve which is shown in Figure 3.11 together with the 95 % confidence bands and the confounder adjusted data points. Figure 3.12 shows the resulting standardized residuals and their QQ plot. The left tail of the distribution shows some deviations from the diagonal indicating normality. Nonetheless, the assumption of normality seems reasonable.

Figure 3.11: Best Fit Model using Fractional Polynomials with Log-Transformed Exposure as Predictor, 95 % Confidence Band and Confounder Adjusted Data

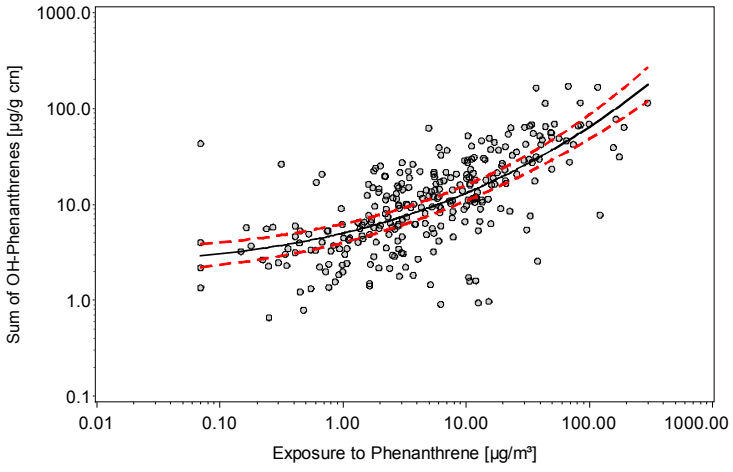
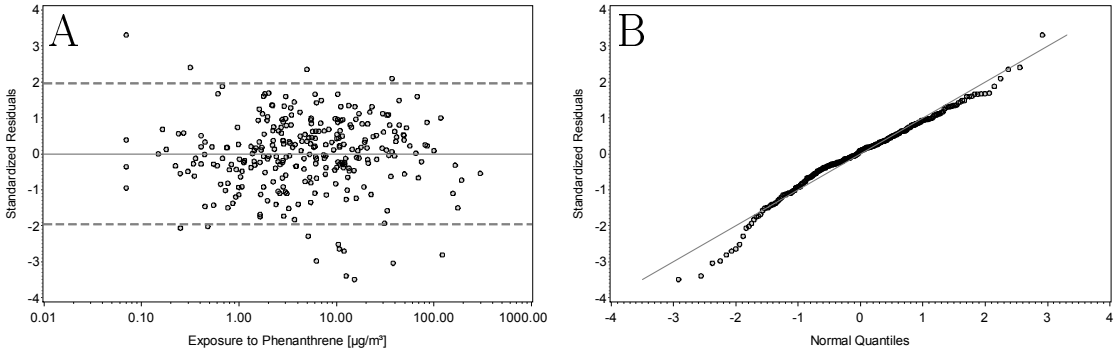


Figure 3.12: Standardized Residuals (A) and QQ Plot (B) of the Best Fit Model using Fractional Polynomials with Log-Transformed Exposure as Predictor



3.2.5 Additive Models

An additive model was applied to explain the excretion of OHPHE with the log-transformed predictor exposure to PHE. For presentation purposes, the results of the smoothing spline are split into a linear component and the deviation from linearity. The linear component can be interpreted as the overall underlying trend of the dose-response curve. The model selection procedure as described in section 2.4 was applied to decide on the necessary amount of flexibility with which the non-parametric part of the model should be provided.

The results of the model selection procedure are given in Table 3.15. The first step of the procedure corresponds to a linear regression which is equivalent to an additive model with a smoothing spline of one degree of freedom (DF). Detailed information on linear regression results can be found in section 3.2.2. The model with a smoothing spline of 2 DF was tested against the first step model. The F-test was highly significant ($P = 0.002$) revealing a better model fit of the additive model with a smoothing spline of 2 DF compared to linear regression. Hence, the model selection procedure was pursued and the model with smoothing spline of 3 DF was tested against the 2 DF model. The corresponding F-test shortly failed significance ($P = 0.052$) and thus, the 3 DF model was not preferred over the 2 DF model. Nevertheless, the model selection procedure was continued for a further step and the additive model with a smoothing spline of 4 DF was calculated. The comparison of this model against the 3 DF model was again not significant ($P = 0.09$). Consequently, the additive model with a smoothing spline of 2 DF was selected as the best fit model.

The model selection procedure based on the F-test can be checked by considering the information criteria and derived quantities presented in Table 3.16. The variance explained by the models increased from the model with smoothing spline of 1 DF (linear regression) to that of 4 DF from 49.1 % to 52.1 %. The model with the best model fit and hence the greatest support suggested by AICC was that with a smoothing spline of 4 DF. Using the AICC differences, the linear regression model was interpreted as having *essentially no support*

Table 3.15: Model Selection Procedure for Additive Models

			Test[a]		
Model	**RSS**[b]	$\mathbf{DF_{Model}}$[c]	**F**	**DF**[c]	***P***
Smoothing spline, 1 DF[d]	194.1	277			
Smoothing spline, 2 DF	187.2	276	10.2	1	0.002
Smoothing spline, 3 DF	184.7	275	3.8	1	0.052
Smoothing spline, 4 DF	182.7	274	2.9	1	0.09

[a]F-test versus previous model; [b]Residual sum of squares; [c]Degrees of freedom; [d]Equivalent to linear regression

($\Delta = 10.7$). The model with 2 DF showed *less support* ($\Delta = 2.6$), whereas the model with 3 DF had *substantial support* ($\Delta = 0.8$). In terms of differences in BIC and the corresponding a-posteriori probabilities, the model with a smoothing spline of 2 DF was identified as the best fit model with an a-posteriori probability of 61.7 %. The other models showed a-posteriori probabilities of 6.2 % for linear regression, 25.5 % for a smoothing spline of 3 DF and 6.6 % for the 4 DF model.

Table 3.16: Model Fit and Information Criteria for the Applied Additive Models

Model	$\boldsymbol{p}$[a]	$\mathbf{R^2}$[b]	**AICC**[c]	Δ_i[d]	**BIC**[e]	$\mathbf{\Delta BIC}_i$[f]	$\boldsymbol{P}_{\mathbf{post}}$[g]
Smoothing spline, 1 DF[h]	1	49.1 %	712.4	10.7	737.5	4.6	6.2 %
Smoothing spline, 2 DF	2	50.9 %	704.3	2.6	733.0	0.0	61.7 %
Smoothing spline, 3 DF	3	51.6 %	702.5	0.8	734.7	1.8	25.5 %
Smoothing spline, 4 DF	4	52.1 %	701.7	0.0	737.4	4.5	6.6 %

[a]Number of model parameters used for the exposure variable; [b]Coefficient of determination; [c]Akaike information criterion (corrected); [d]$\Delta_i = \mathrm{AICC}_i - \mathrm{AICC}_{\min}$; [e]Bayesian information criterion; [f]$\Delta\mathrm{BIC}_i = \mathrm{BIC}_i - \mathrm{BIC}_{\min}$; [g]a-posteriori probability; [h]Equivalent to linear regression

Parameter estimates of the selected model are given in Table 3.17. The level of OHPHE under an exposure to PHE of 1 μg/m^3 (intercept) was 4.95 μg/g crn (95 % CI 3.94–6.22 μg/g crn). The transformed regression parameter of the linear component was 1.39 with a 95 % CI of 1.33–1.45 ($P < 0.0001$). This means, in an overall trend, OHPHE increased by 39 % under a doubling of the exposure to PHE. Obviously, no estimates can be given for the non-parametric part of the model. Thus, for this part only a test for deviations of linearity is presented which reveals a significant better model fit that is obtained by including the non-parametric part ($P = 0.001$).

The estimated dose-response curve is illustrated in Figure 3.13 together with 95 % confidence bands and confounder adjusted data. It can be seen that deviations from linearity are present especially in case of lower values of PHE. Figure 3.14 shows the resulting standardized

residuals and the corresponding QQ plot for the check of a normal distribution. As in case of the other models, the left tail of the distribution showed some deviations from normality. However, overall the assumption of normality seems reasonable.

Table 3.17: Results of the Additive Model with Log-Transformed Exposure to PHE

Variable	**DF**[a]	**exp(β)**[b]	**95 % CI**[c]	***P***
Intercept	1	4.95	$(3.94, 6.22)$	<0.0001
$\log_2(\text{PHE})$	1	1.39	$(1.33, 1.45)$	<0.0001
$\text{Spline}\big(\log_2(\text{PHE})\big)$	1	–	–	0.001

[a]Degrees of freedom; [b]Back-transformed parameter estimate adjusted for type of industry and current smoking; [c]Confidence interval of exp(β)

Figure 3.13: Best Fit Model using Additive Models with Log-Transformed Exposure as Predictor, 95 % Confidence Band and Confounder Adjusted Data

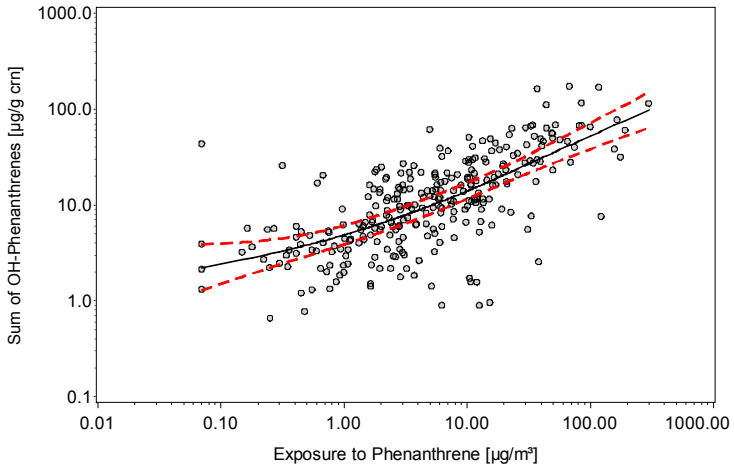
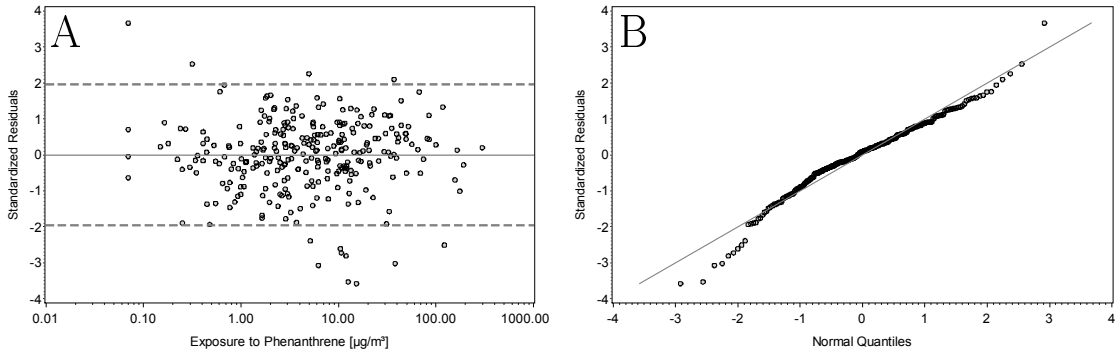


Figure 3.14: Standardized Residuals (A) and QQ Plot (B) of the Best Fit Model using Additive Models with Log-Transformed Exposure as Predictor



3.2.6 Model Comparison

The above selected models for linear splines, fractional polynomials, additive models as well as ANOVA and linear regression are compared by means of the information criteria and derived quantities presented in section 2.6. The results of these calculations are given in Table 3.18. The best model fit for each criterion is underlined.

The model complexity of the selected models, reflected by the number of model parameters used for the exposure variable, varied by 3 with a minimum of 1 for linear regression and a maximum of 4 for fractional polynomials with untransformed exposure variable. The proportion of variation explained by the selected models differed by 2.6 percentage points (minimum: linear regression with 49.1 %, maximum: linear splines 51.7 %).

The minimum AICC of the selected models was achieved by linear splines with one knot and hence revealed the greatest support. ANOVA and linear regression showed AICC differences to the linear splines model of $\Delta = 14.4$ and $\Delta = 10.4$ indicating *essentially no support* for the two models. Fractional polynomials with untransformed and log-transformed exposure variable as well as additive models showed values of Δ between 2.3 and 3.3 indicating *less support*.

The best model fit using BIC was detected for the additive model. Calculations of the a-posteriori probabilities over the BIC differences exhibited a probability of 43.7 % for this model. Linear splines and fractional polynomials with log-transformed exposure variable presented a-posteriori probabilities of 24.1 % and 26.6 %. ANOVA, linear regression and fractional polynomials on the untransformed exposure variable showed least support with a-posteriori probabilities of 0.02 %, 4.4 % and 1.2 %, respectively.

Table 3.18: Model Fit and Information Criteria for the Applied Methods

Model	p[a]	$\mathbf{R^2}$[b]	**AICC**[c]	Δ_i[d]	**BIC**[e]	Δ**BIC**$_i$[f]	$P_{\mathbf{post}}$[g]
ANOVA	3	49.2 %	716.4	14.4	748.6	15.6	0.02 %
Linear Regression	1	49.1 %	712.4	10.4	737.5	4.6	4.4 %
Linear Splines	3	<u>51.7 %</u>	<u>702.0</u>	0.0	734.1	1.2	24.1 %
Fractional Polynomials on PHE	4	51.7 %	704.4	2.5	740.1	7.2	1.2 %
Fractional Polynomials on $\log_2$(PHE)	2	50.8 %	705.3	3.3	734.0	1.0	26.6 %
Additive Model	2	50.9 %	704.3	2.3	<u>733.0</u>	0.0	43.7 %

[a]Number of model parameters used for the exposure variable; [b]Coefficient of determination; [c]Akaike information criterion (corrected); [d]Δ_i = AICC$_i$ − AICC$_{\min}$; [e]Bayesian information criterion; [f]ΔBIC$_i$ = BIC$_i$ − BIC$_{\min}$; [g]a-posteriori probability

4 Discussion

Analysis of dose-response curves is a frequently used instrument to understand physiological mechanisms and dependencies of analyzed variables. Estimates of the functional relation between predictor and outcome play an important role in risk assessment, e. g. in the regulatory process of occupational medicine for defining limits of exposure in the workplace. The US Environmental Protection Agency (EPA) has developed several guidelines for risk assessment of carcinogenic substances (EPA, 1996). Standard methods are for example, cumulative exposure measures and the assumption of a linear dose-response relation. However, it turned out that even for well examined exposures an optimal exposure metric often cannot be formulated. Consequently a uniform modeling approach for dose-response relations does not exist either.

Linear Regression

A common simple approach for statistical analysis of dose-response curves is the assumption of a linear relation between the variables of interest and the application of standard linear regression methods to calculate estimates of the model parameters. This method is well established and easy to communicate. While this assumption often seems reasonable at first glance, many settings might lead to a transgression of linearity in practice. For example, thresholds are suggested for some substances leading to an abrupt change of the dose-response curve. In the low-dose range other effects such as sensitization or hormesis are discussed, e. g. protective by induction of enzymes. For high doses, overload or saturation might lead to a weakened relation between predictor and outcome. In occupational settings, this effect might be observed due to the existence of a healthy-worker effect. The occurrence of one or more of these effects might consequently lead to quite a complex shape of the dose-response curve.

Analysis of Variance

Another approach widely applied to model dose-response curves is ANOVA. Therefore, the continuous predictor is transformed into a class variable. While a constant level of the outcome is claimed within the categories, no assumptions concerning the shape of the underlying dose-response curve between the categories have to be made. This approach often serves as an alternative to linear regression because of the more flexible form the estimated relation can

take. Similar to linear regression, ANOVA offers the benefit of an easy interpretation and communication of results. However, the classification of the continuous predictor causes an important loss of information and is criticized by many authors (Greenland, 1995b; Thurston, Eisen and Schwartz, 2002; Steenland and Deddens, 2004; Altman and Royston, 2006). In general, the assumption of a constant level of the outcome within the categories of the predictor seems unrealistic for most situations. In addition, the choice of the category boundaries can rarely be based on a sound rationale. User-driven categories often seem quite arbitrary, whereas optimal category bounds may overestimate the true relation. Another criticism is the risk of differential misclassification as a result of measurement error of the predictor and misclassification in the categories.

Fractional Polynomials

In the mid nineties, Royston and Altman (1994) introduced fractional polynomials as an alternative to polynomial regression. The predictor variable is modeled using a relatively small set of transformations. The authors pointed out that the set of functional forms that can be modeled by fractional polynomials is sufficiently flexible to cope with most of the problems in the real world. In particular, the set comprises J-shaped functions that are discussed for effects of *hormesis* at lower exposure levels. A detailed discussion on fractional polynomials is given in Royston and Altman (1994).

In order to decide on which transformations of the set of possible transformation of the independent variable should be applied in the final model, Royston and Altman propose a stepwise approach. This is criticized by Greenland (1995a), who suggests a more intuitive approach under consideration of prior knowledge of the characteristic of the dose-response curve. In spite of this criticism, here the original stepwise approach was applied in order not to influence the shape of the dose-response curve by subjective decisions as well as to ensure an easier comparison with the other applied methods.

A problem for the application of fractional polynomials are variables with negative or zero values. These values cannot be included in the analysis because the application of the logarithm or the square root would produce missing values. Royston and Altman propose adding a constant to these variables to eliminate non-positive values. However, this introduces an additional parameter into the model and it is not clear how the constant should be chosen. Royston and Sauerbrei (in press) propose the use of a transformation that can be applied both to shift the origin away from zero and to weigh down the influence of values with high leverage.

As the dataset of the PAH study only shows positive values for the exposure to PHE, this problem did not occur by applying fractional polynomials on the untransformed exposure variable. The proposed transformation of Royston and Sauerbrei could have been applied

anyhow to weigh down values with high leverage that were present at high exposures. However, to be in consistence with the application of the other methods and to provide easy comparisons between the methods, this was not done.

As regards the use of the log-transformed exposure to PHE, values below $1\,\mu g/m^3$ resulted in non-positive values of the predictor variable. This could easily be solved by multiplying the original exposure variable by 1000, i. e. expressing the exposure in the unit ng/m^3. Again, in order to be consistent with the other methods, it was decided not to apply the transformation of Royston and Sauerbrei.

In spite of fractional polynomials being a parametric method, the parameters in general are poorly or not interpretable. The method therefore is located between parametric and non-parametric regression (Greenland, 1995a) and the fitted dose-response curve has to be interpreted largely in a graphical manner.

Additive Models

Additive models and the larger class of generalized additive models were introduced by Hastie and Tibshirani (1986) and are a generalization of the (generalized) linear regression model by addition of a non-parametric smoother into the linear model. Depending on the amount of flexibility of the non-parametric part expressed in terms of equivalent degrees of freedom, additive models are able to detect a large set of underlying functionals. A detailed discussion of the method can be found in the work of Hastie and Tibshirani.

Despite being a generalization of the linear model, additive models in general serve a different analytic purpose. While the emphasis by using linear regression is mainly on parameter estimation and statistical inference, additive models represent a more explorative approach through a visualization of the relation between outcome and predictor. However, if semi-parametric models are applied, parameter estimates and tests of this part of the model may be interpreted in a similar way as in linear regression.

As for most non-parametric regression techniques, a parameter of smoothness has to be chosen which for additive models consists of the choice of the equivalent degrees of freedom. This issue can be compared to the choice of the number of cut points for categorical analyses or linear splines and the choice on the transformation for fractional polynomials. Here, the number of equivalent degrees of freedom was chosen by increasing the degrees of freedom one by one and comparing each model fit with the preceding one.

As described in section 2.4, each predictor that is modeled non-parametrically is divided into a linear part and a non-parametric deviation from linearity. The test for the linear part searches for an overall linear trend in the relation, whereas the test for the non-parametric part tries to detect deviations from overall linearity. The tests for both parts can be interpreted

together and can provide suitable information about the underlying relation between predictor and outcome.

Linear Splines

Splines are piecewise polynomials that join in a smooth way at the so-called knots and are a flexible way of fitting an unknown functional form (de Boor, 1978). They are a powerful method suitable for describing a large class of functional forms. For most applications of splines, cubic splines are chosen, i. e. polynomials of degree three between two subsequent knots (Brown, Ibrahim and DeGruttola, 2005). For example Hauptmann et al. (2002) used cubic splines to model the weights of a cumulative exposure measure. However, an inconvenience of this method is that no interpretable parameters are provided (Steenland and Deddens, 2004). Theoretically, a knot is a point where the spline alters its characteristic and hence this point might be of special interest for the interpretation of the fitted function. Nevertheless, as this change in characteristic only becomes evident at the third derivative, in practice the dose-response curve might exhibit no specific feature at the knot location. Moreover, the remaining parameters corresponding to the spline basis functions cannot be interpreted alone either, regardless of the basis used. Therefore, cubic splines are closer to non-parametric than parametric methods (Greenland, 1995a) and − similar to additive models and fractional polynomials − are of a more explorative nature. Principally, important characteristics of the dose-response curve, e. g. thresholds, are derived graphically by analysis of the chart of the fitted function. However, most cases require more than a graphical representation of a dose-response curve, i. e. a measure of the strength of the relation between predictor and outcome as well as tests or confidence intervals.

Linear splines can be a promising approach to meet these expectations without the mentioned inconvenience of non-interpretable parameters. They are piecewise linear functions, which are connected at the knots (Molinari, Daures and Durand, 2001; Rosenberg et al., 2003). In this way, the feature of smoothness of the spline function reduces to continuity. But, depending on the number of knots, linear splines can exhibit a large amount of flexibility. Although recommended in Greenland (1995a) as an alternative to standard dose-response and trend analysis, their application is more uncommon than that of cubic splines. Only a few papers using linear splines have been published in the last few years (Molinari, Daures and Durand, 2001; Muggeo, 2003; Li and Hunt, 2004; Bessaoud, Daures and Molinari, 2005).

The use of linear splines is highly associated to ANOVA with a classified linear predictor. In fact, the latter can be seen as splines of degree $d = 0$ or constant splines. Within the classes constituted by the knots, the assumption no longer consists of a constant level of the outcome with discontinuities at the knots as for ANOVA, but of a constant linear influence of the predictor on the outcome. Thus, the classes are no longer compared by outcome levels,

but by slopes of the regression line. The problem of misclassification into the classes due to measurement errors is reduced because of the absence of discontinuities (Greenland, 1995a).

The application of linear splines leads to a simple parametric model as recommended by Steenland and Deddens (2004), because a knot can be interpreted as a threshold where the slope of the dose-response function changes. A direct interpretation of the remaining parameters depends on the basis used for the space of linear splines. In case of using the B-spline basis, this is not straightforward and in general not of great concern. The parameters of the truncated power basis functions represent the change of the slope from one segment to another, i. e. a parameter estimate of 0 corresponds to no change in slope of the fitted function at the corresponding knot location. Finally, by use of the double truncated linear basis (DTL) that was developed and applied in this work, the parameters correspond to the slope of the fitted function in the corresponding segment. This implicates the advantage of a parameter interpretation as is usual in linear regression.[1]

A methodological problem for the application of splines is the choice of the number as well as the location of the knots. In most cases, only a few well-placed knots will suffice to model the relationship between variables (Molinari, Daures and Durand, 2001). If the knots are known a priori, standard regression methods can be applied. Nevertheless, in most situations this is not the case and a reasonable choice of the locations and numbers of knots have to be made. Some recommendations are given in the literature. In general, two approaches can be distinguished: preselection of the knot locations by the user and automatic methods. The user's choice is made either by visual inspection or more objectively, by using quantiles as knot locations. The automatic methods seek to optimize a goodness-of-fit criterion.

Rosenberg et al. recommend the use of quantiles for the knot locations and an increasing number of knots, while the best fit model is chosen by means of the Akaike Information Criterion (AIC) (Rosenberg et al., 2003). This approach assures an adequate amount of flexibility for a cubic spline and is chosen for example in the work of Hauptmann et al. (2002). However, for linear splines, where the knots can be interpreted directly as points of change of the functional relationship between the studied variables, this approach may miss important knot locations. Meanwhile, it may be necessary to test if such change points exist and to make inference about their location, which is clearly not possible with preselected knots.

In order to do this, an automatic selection of the knots has to be made. Molinari, Daures and Durand (2001) maximized the partial likelihood with respect to the knots. Inference about the knot locations were drawn by bootstraps. However, this procedure can invalidate conventional tests and confidence intervals on the model parameters (Greenland, 1995a).

[1] Nevertheless, it has to be stated that slope estimates for the segments can also be calculated by using B-splines or the truncated power basis by means of linear combinations of the regression parameters. The difference between the three bases is only apparent for the direct interpretation of the parameters without any transformations.

Another possibility that was chosen in this work, is to consider knots directly as model parameters and to fit the model as a whole. Consequently, the resulting model is no longer linear in the parameters but becomes a non-linear model. Inference on the knot locations and the other model parameters follow the usual non-linear regression theory.

A special feature of automatic knot selection with linear splines in contrast to higher order splines is the risk of a local overfit of the data. As continuity is the only smoothness criterion of linear splines, the characteristic of the spline function can change dramatically at a knot location. If a segment formed by two knots becomes too small, i. e. contains too few data points, the regression line in this segment might become unstable. This is especially the case for the two segments at the edges and for two small subsequent interior segments. In the extreme case, the fitted spline function might try to interpolate outliers in these segments. A strategy to reduce the risk of local overfitting is given by defining a lower limit on the number of observations included in a segment. In this work, this was realized by ensuring that each segment had at least 10 % of the total observations, i. e. $N_{\text{seg}} \geq 29$.

PAH Study

Biomonitoring studies should be scientifically robust to provide sound estimates for potential risks to human health (Bates et al., 2005). Personal measurements of the exposure are considered superior to stationary sampling (Kromhout and van Tongeren, 2003), and models based on individual exposure measurements are superior to categorical analyses based on exposure groups (Greenland, 1995b). Reasonable simple models should be applied that yield biologically interpretable parameters (Steenland and Deddens, 2004).

Statistical analyses were carried out to characterize the dose-response relation for exposure to PAHs and excretion of urinary metabolites among a large group of German workers. Overall, there was a strong association between external and internal exposure where PHE explained about 50 % of the variance of OHPHE.

This biomonitoring study was conducted in occupational settings with high exposure to PAHs such as coke production and the manufacturing of refractory products and graphite electrodes. Exposure to 16 PAH compounds were determined with personal measurements during a working shift. The distributions of all substances were highly skewed. The lower quartile of the B[a]P concentrations was 86 ng/m^3 and exceeded typical environmental settings where levels of less than 10 ng/m^3 have been measured (Straif et al., 2005). The upper quartile of 1420 ng/m^3 B[a]P was higher than in occupational settings with a few exceptions such as the top side of coke ovens or in pot rooms of aluminum smelters where exposure can be as high as 100 $\mu\text{g/m}^3$ (Armstrong et al., 2004). The large fraction of measurements below the LQ (23.9 %) rendered B[a]P and other carcinogenic PAHs less suitable for dose-response modeling. Among the various monohydroxylated metabolites of PAH detected in human urine

(Grimmer et al., 1990), OHPHE and 1-OH-pyrene were determined in the present study. As observed for external exposure, internal exposure levels were also highly skewed. The 5^{th} percentile of occupational 1-OH-pyrene levels corresponded to the 95^{th} percentile in U.S. males investigated in the 1999-2000 National Health and Nutrition Examination Survey (CDC, 2005). On average, the workers showed 30-fold higher levels of urinary 1-OH-pyrene than smokers in the general German population (Becker et al., 2003). For OHPHE, the levels were about 10-fold higher than in German smokers (Umweltbundesamt, 1998). OHPHE levels were less affected by smoking and type of industry than 1-OH-pyrene (Rihs et al., 2005).

The more general question remains, which suitable and sufficient biomarkers of occupational or environmental PAH exposure should be determined in future studies. Although B[a]P is considered a better candidate than PHE for monitoring exposure to PAHs due to its carcinogenicity, its tetrol is difficult to determine (Simpson et al., 2000) and excreted in urine only in tiny amounts, even at high exposure levels (Wu et al., 2002). However, as phenanthrene is the simplest PAH with a bay region, a structural feature associated with carcinogenic properties of higher molecular weight PAHs such as B[a]P, it can serve as a surrogate compound that mirrors the metabolic activation of carcinogenic PAHs to diol epoxides (Hecht et al., 2003).

In order to characterize the shape of a dose-response relation between exposure to PHE in the workplace and the urinary excretion of OHPHE, ANOVA, linear regression, linear splines, fractional polynomials and additive models were applied and compared. Current smoking and type of industry were included in all models as independent variables to control for confounders. Current smoking had no impact on the excretion of OHPHE, while type of industry turned out to have a significant influence on the excretion of OHPHE ($p = 0.001$). This might be explained by the different settings in the industries, different composition of the PAH exposure and unmeasured exposure routes. A few workers used protective equipment. Measurements not taken behind the mask may have contributed to a weakened association between external and internal exposure. Dermal exposure can be another important route of PAH exposure in certain occupational settings (Boogaard and van Sittert, 1995; McClean et al., 2004). In the study at hand, dermal exposure was not assessed, but protective clothing was used especially in occupational settings with relevant dermal exposure.

The different models explained about 49 to 52 % of the variance of OHPHE. The unexplained variance of the association between external and internal exposure may have several causes. General reasons for exposure variability and its effect on exposure assessment have been reviewed (Loomis and Kromhout, 2004; Lin, Kupper and Rappaport, 2005). Sufficient information on the measurement strategy should be provided to estimate the uncertainties of the exposure variable. In this study, external exposure was assessed with personal measurements in the worker's breathing zone during a working shift. However, measurements were only taken for two hours on average. Exposure variability during the working shift may have led to exposure misclassification.

Another limitation of the study at hand is the assessment of internal exposure with spot urine samples because the collection of 24 hour urine samples was less feasible. Also the error of OHPHE measurements can contribute to the unexplained variance (Carmella et al., 2004). Further, insufficiently controlled confounders may add to the residual confounding. Smoking was implemented only as a categorical variable (current smokers and non-smokers) because o-cotinine measurements were not available for all workers. However, models with o-cotinine concentrations did not significantly improve the model fit. For the majority of workers, the occupational exposure levels were much higher than smoking-related PAH exposure. OHPHE showed a stronger correlation with PHE exposure than 1-OH-pyrene with pyrene and was less affected than 1-OH-pyrene by smoking and type of industry (Rihs et al., 2005). Other potential confounders such as age and Caucasian ethnicity had no relevant impact on OHPHE in this study (Rihs et al., 2005). Also genetic polymorphisms may modulate the individual levels of PAH metabolites (Wu et al., 1998; Alexandrie et al., 2000; Nerurkar et al., 2000; Kuljukka-Rabb et al., 2002; Wu et al., 2002; Kim et al., 2003; Rihs et al., 2005).

The results of ANOVA and linear regression are presented in sections 3.2.1 and 3.2.2. Linear regression showed a highly significant relation between the exposure to PHE in the workplace and the urinary excretion of OHPHE ($P < 0.0001$). Under a doubling of the exposure to PHE, OHPHE increased by about 40 %. The analysis with the categorized exposure variable also showed a highly significant relation between PHE and OHPHE ($P < 0.0001$) and a monotonic increase of the levels of OHPHE with increasing exposure categories.

By applying linear splines for the analysis of the relation between PHE and OHPHE, one knot was identified at 0.77 μg/m^3 PHE which corresponded to a estimated internal exposure of about 3 μg/g crn OHPHE. This value was above the 95$^{\text{th}}$ percentile in the general German population (Umweltbundesamt, 1998), whereas in the PAH study only 13 % of workers had lower levels of internal exposure. In the lower dose segment, there was no association between external and internal exposure. In the upper dose segment, there was a clear linear increase of internal exposure by external exposure at the log-transformed scales.

Fractional polynomials with the use of the untransformed exposure variable showed similar results to linear splines. A two degree model was chosen as the best fit with -1 and 0 as exponents, i. e. x^{-1} and $log_2(x)$ as transformations of the exposure to PHE. For the high-dose range (about $\geq 1\,\mu g/m^3$), the influence of x^{-1} vanishes and the log-transformation becomes dominating. For this region, OHPHE increases by a factor of 1.48 under a doubling of the exposure to PHE ($P < 0.0001$). For the low-dose range (about $\leq 0.2\,\mu g/m^3$), the estimate of the dose-response curve shows a decrease of OHPHE under an increase of PHE. A test for this decrease is not feasible, thus the 95 % confidence band indicates that a constant level of the outcome in this region seems possible too.

Fractional polynomials with the log-transformed exposure variable resulted in a model of degree 1 with a transformation of the exposure variable by x^3. The resulting estimate of the

dose-response curve shows for the low-dose range an increase of OHPHE by a factor of 1.21 under a doubling of PHE from 0.5 to 1 μg/m^3. In the high-dose range OHPHE increased by a factor of 1.74 under a doubling of PHE from 50 to 100 μg/m^3.

The results of the additive model are comparable to that of the fractional polynomials using the log-transformed exposure variable but less pronounced. A model with 2 degrees of freedom for the splines component was identified as the best fit model. The model reveals an overall linear trend with an increase of OHPHE by a factor of 1.39 under a doubling of PHE. The deviation of linearity leads to an increase of OHPHE by a factor of 1.27 under a doubling of PHE from 0.5 to 1 μg/m^3 in the low-dose range and to an increase by a factor of 1.49 under a doubling of PHE from 50 to 100 μg/m^3 in the high-dose range.

Overall, a highly significant association between the exposure to PHE and the excretion of OHPHE was identified by each of the methods. With the exception of linear regression, the dose-response estimates showed a different shape in the low-dose range compared to the high-dose range with higher increases of OHPHE at higher doses of PHE. Linear splines and fractional polynomials with the untransformed exposure even showed a slight decrease of OHPHE in the low-dose range. An estimate of the boundary of these two regions is available by using linear splines via the knot estimate. However, the location of the knot is data driven, and interpretation should be performed with caution (Ulm and Salanti, 2003). One way of interpreting a knot is mechanistically, as a toxicokinetic threshold (Molinari, Daures and Durand, 2001). Such thresholds are discussed for several agents (Bolt, 2003; Popp, Crouch and McConnell, 2005). In particular the low-dose range is a matter of debate (Calabrese and Baldwin, 2000; Thayer et al., 2005). Furthermore, methodological issues should be considered regarding a higher uncertainty of data in the low-dose range. PAHs from non-occupational sources may confound the association especially at lower doses, whereas occupational exposure at high doses may override confounding with smoking or diet. What is more, the margins of the dose-response curve are a specific problem when using splines. Therefore, artifacts also have to be taken into account when a knot is located in the margin of the dose range.

For the high-dose range, the factor of the increase of OHPHE under a doubling of PHE varied from 1.38 to 1.74 between the applied models. Thus, the increase in excretion of OHPHE in urine is relatively slower under an increase of external exposure to PHE. That raises the question of whether competing metabolic pathways other than the detoxification of PHE to OHPHE are more preferred at higher exposure levels and if so, which other urinary metabolites of PHE should be considered. PAHs represent complex mixtures. Their chemistry and formation has been reviewed by the International Agency for Research on Cancer (IARC, 1983; IARC, 2006). Of particular interest is the metabolic activation pathway of PHE where phenanthrene 1,2-dihydrodiol (Jacob, Grimmer and Dettbarn, 1999) and r-1,t-2,3,c-4-tetrahydroxy-1,2,3,4-tetrahydrophenanthrene (PheT) (Hecht et al., 2003) can be determined as urinary biomarkers. Exposure to PHE such as by smoking seems to induce the diol epoxide

pathway of PHE (Hecht et al., 2005). Although the metabolism of PAHs is among the most extensively studied pathways, it is complex and yet to be fully elucidated.

Regression techniques always need to be coupled with diagnostics to inspect the model fit and to identify influential data points (Greenland, 1995a). This can be achieved by inspection of the standardized residuals as well as QQ plots. The characteristics of these plots are very similar for the different models, with two groups of observations especially showing conspicuous results. One of these groups is composed of workers with a considerable exposure to PHE of about $10\,\mu g/m^3$ and a relatively low level of OHPHE of about $1\,\mu g/g\,crn$. All models fail to fit these observations. Furthermore, in the QQ plots they can be found among the leftmost observations representing the largest negative residuals. However, the influence of the observations on the estimated dose-response curve is relatively small due to the high number of observations with a similar exposure and higher levels of OHPHE. The other noticeable group of observations is formed by workers exhibiting a low exposure to PHE ($\leq 1\,\mu g/m^3$) and at the same time relatively high levels of OHPHE ($\geq 10\,\mu g/g\,crn$). One worker especially exhibits quite unusual values with an exposure to PHE below the limit of quantification ($0.07\,\mu g/m^3$) and an excretion of OHPHE of about $48\,\mu g/g\,crn$ which approximately corresponds to the 90 % quantile. The observations were checked for correctness, but no measurement or typing error could be identified. In order to detect the influence of the data points on the estimates of the dose-response curve, the analyses were applied a second time without the respective observations. The results of the models changed only slightly and all test decisions during the model selection procedure for the different methods remained unchanged.

Model comparison by AICC and BIC and derived information revealed two groups among the applied methods. Fractional polynomials, additive models and linear splines showed a considerable better model fit than ANOVA and linear regression for both criteria. The best model fit assessed by AICC was achieved by linear splines. Following the recommendations of Burnham and Anderson (2004) for the interpretation of the differences in terms of AICC, ANOVA and linear regression showed *essentially no support* in comparison to linear splines. Fractional polynomials and additive models were close to linear splines but failed to reach the category of *substantial support*. The best model fit assessed by BIC, which tends to prefer simpler models, was realized by the additive model exhibiting only two parameters for modeling the association between PHE and OHPHE. Assuming equal a-priori probabilities for all models, the additive model had an a-posteriori probability of 43.7 %. Considerable a-posteriori probabilities were achieved as well by fractional polynomials on the log-transformed exposure (26.6 %) and by linear splines (24.1 %). The other models showed a-posteriori probabilities below 5 %. Taking into account the number of parameters, they can be considered as either exhibiting too poor a model fit (linear regression), being too complex (fractional polynomials using the untransformed exposure) or both (ANOVA). Altogether, the linear splines model yielded the best compromise between model complexity and model fit.

5 Conclusions

Greenland (1995a) states that analyses of dose-response and methods to control for continuous confounders should not be restricted to categorical, linear or trend test approaches. This can be easily provided by fractional polynomials or spline regression. The author encourages the use of these methods and mentions their easy implementation in standard analysis packages.

Steenland and Deddens (2004) request simple parametric models for dose-response modeling due to three reasons: (1) the underlying relation between exposure and response is usually of a simple and monotonic form, (2) simple models are easier to communicate and will be used more frequently by subsequent users and (3) they are the best tool for regulatory affairs.

Linear splines are an example of such a simple parametric model for dose-response analyses. A major benefit of their use is the good interpretability of the model parameters. The knots divide the continuous exposure variable into segments of potentially different influence on the outcome. Slope estimates can be interpreted directly as the influence of the exposure variable on the outcome in the corresponding segment. All model parameters with exception of the knots enter linearly into the model, i. e. if knots are considered as fixed, standard linear regression methods can be applied. If the knots are considered as model parameters, a non-linear model can be used and the non-linear theory for estimation and tests applies. Information on the knots can be used to determine regions of interest of the dose-response curve or to derive thresholds, e. g. to determine a maximum tolerable exposure concentration in the workplace[1].

The use of linear splines is not restricted to modeling continuous outcomes. They can also be applied in logistic regression, proportional odds models or survival analysis. They show a great flexibility for fitting a large class of underlying relations. They provide a simple model and – together with the DTL basis – easily interpretable parameters without the necessity of previous transformations. The model selection procedure proposed within this work offers the possibility of detecting the number of knots necessary to describe the underlying dose-response curve, while the limitation on the number of data points between two knots limits the risk of overfitting. Overall, linear splines are a suitable approach to describe dose-response relations and can be considered as a compromise between the standard techniques as ANOVA and linear regression and the more complex methods such as fractional polynomials and additive models.

[1] German: Maximale Arbeitsplatzkonzentration (MAK)

Summary

The development of an appropriate model plays an important role for the estimation of unknown functional relations between medical, biological or epidemiological parameters. Such models can provide insight in the underlying mechanisms and be a basis during regulatory processes. However, the existing standard methods for analyzing the influence of a continuous predictor, such as analysis of variance or linear regression, exhibit numerous causes for criticism. The aim of this work is to examine linear splines for modeling dose-response relations in comparison to these standard methods, as well as to the more complex techniques of fractional polynomials and additive models. The methods are applied and compared to a dataset from an occupational study that examines the effects of exposure to polycyclic aromatic hydrocarbons in the workplace. In this context, the dose-response relation between exposure to phenanthrene and excretion of the urinary metabolites 1-, 2-+9-, 3- and 4-OH-phenanthrene is analyzed. Linear Splines, fractional polynomials and additive models are superior to the standard methods regarding the model fit. All three methods show a non-existant or weak relation between external and internal exposure in the low-dose range, while a clear influence becomes apparent in the high-dose range. Additionally, linear splines yield an estimate for the boundary between the two regions. Overall, the use of linear splines leads to a simple parametric model that is easy to communicate and present. Meanwhile it remains sufficiently flexible to fit complex shapes of dose-response curves. Linear splines represent a good compromise between standard methods and more complicated non-linear or non-parametric methods.

Zusammenfassung

Die Entwicklung eines geeigneten Modells zur Schätzung von unbekannten funktionalen Zusammenhängen zwischen medizinischen, biologischen oder epidemiologischen Parametern spielt eine wichtige Rolle zur Gewinnung von Einblicken in zugrunde liegende Mechanismen und im Rahmen von regulatorischen Prozessen. Die existierenden Standardverfahren zur Analyse solcher Beziehungen wie Varianzanalyse oder lineare Regression bieten jedoch viel Anlass zu Kritik. Ziel dieser Arbeit ist die Untersuchung von linearen Splines zur Modellierung von Dosis-Wirkungs-Beziehungen im Vergleich zu diesen Standardverfahren sowie zu den komplexeren Verfahren fractional polynomials und additive Modelle. Die Methoden werden angewandt und verglichen anhand eines Datensatzes einer arbeitsmedizinischen Studie zur Untersuchung der Effekte von beruflicher Exposition gegenüber polyzyklischen aromatischen Kohlenwasserstoffen. In diesem Zusammenhang wird die Dosis-Wirkungs-Beziehung zwischen Exposition gegenüber Phenanthren und der Ausscheidung der Metabolite 1-, 2-+9-, 3- und 4-OH-Phenanthren im Urin analysiert. Lineare Splines, fractional polynomials und additive Modelle sind den Standardverfahren in Bezug auf die Anpassung der Daten deutlich überlegen. Alle drei Verfahren zeigen im Niedrig-Dosis-Bereich einen nicht existenten oder schwachen Zusammenhang zwischen äußerer und innerer Exposition, während im Hoch-Dosis-Bereich ein klarer Einfluss deutlich ist. Lineare Splines liefern darüber hinaus einen Schätzwert für die Abgrenzung dieser beiden Bereiche. Insgesamt führt die Verwendung von linearen Splines zu einem einfachen, gut kommunizierbaren und leicht darstellbaren parametrischen Modell, das jedoch ausreichend flexibel ist, um auch komplexe Verläufe von Dosis-Wirkungs-Kurven abzubilden. Lineare Splines stellen einen guten Kompromiss dar zwischen den Standardmethoden und komplizierteren nichtlinearen oder nichtparametrischen Methoden.

Bibliography

1. Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csàki, F. C. (Eds.): Second International Symposium on Information Theory. Budapest: Adademia kiadó, 267–281.

2. Alexandrie, A. K., Warholm, M., Carstensen, U., Axmon, A., Hagmar, L., Levin, J. O., Ostman, C. and Rannug, A. (2000): CYP1A1 and GSTM1 polymorphisms affect urinary 1-hydroxypyrene levels after PAH exposure. Carcinogenesis, 21, 669–676.

3. Altman, D. G. and Royston, P. (2006): The cost of dichotomising continuous variables. Brit. Med. J., 332, 1080.

4. Armstrong, B., Hutchinson, E., Unwin, J. and Fletcher, T. (2004): Lung cancer risk after exposure to polycyclic aromatic hydrocarbons: A review and meta-analysis. Environ. Health Persp., 112, 970–978.

5. Bates, M. N., Hamilton, J. W., Lakind, J. S., Langenberg, P., O'Malley, M. and Snodgrass, W. (2005): Workgroup report: Biomonitoring study design, interpretation, and communication – Lessons learned and path forward. Environ. Health Persp., 113, 1615–1621.

6. Becker, K., Schulz, C., Kaus, S., Seiwert, M. and Seifert, B. (2003): German Environmental Survey 1998 (GerES III): Environmental pollutants in the urine of the German population. Int. J. Hyg. Envir. Heal., 206, 15–24.

7. Bessaoud, F., Daures, J. P. and Molinari, M. (2005): Free knot splines for logistic models and threshold selection. Comput. Meth. Prog. Bio., 77, 1–9.

8. Bolt, H. M. (2003): Genotoxicity-threshold or not? Introduction of cases of industrial chemicals. Toxicol. Lett., 140-141, 43–51.

9. Boogaard, P. J. and Sittert, N. J. van (1995): Urinary 1-hydroxypyrene as biomarker of exposure to polycyclic aromatic hydrocarbons in workers in petrochemical industries: Baseline values and dermal uptake. Sci. Total. Environ., 163, 203–209.

10. Boor, C. de (1978): A Practical Guide to Splines. Volume 27, Appl. Math. Sci.. New York, USA: Springer.

11. Bostrom, C. E., Gerde, P., Hanberg, A., Jernstrom, B., Johansson, C., Kyrklund, T., Rannug, A., Tornqvist, M., Victorin, K. and Westerholm, R. (2002): Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. Environ. Health Persp., 110 Suppl 3, 451–488.

12. Breiman, L. and Friedman, J. H. (1985): Estimating optimal transformations for multiple regression and correlation (with discussion). J. Am. Stat. Assoc., 80, 580–619.

13. Brown, E. R., Ibrahim, J. G. and DeGruttola, V. (2005): A flexible B-spline model for multiple longitudinal biomarkers and survival. Biometrics, 61, 64–73.

14. Buja, A., Hastie, T. J. and Tibshirani, R. J. (1989): Linear smoothers and additive models (with discussion). Ann. Stat., 17, 453–555.

15. Burnham, K. P. and Anderson, D. R. (2004): Multimodel inference: Understanding AIC and BIC in model selection. Amsterdam Workshop on Model Selection, Online Publication: `http://www2.fmg.uva.nl/modelselection/presentations/AWMS2004-Burnham-paper.pdf`.

16. Calabrese, E. J. and Baldwin, L. A. (2000): Tales of two similar hypotheses: The rise and fall of chemical and radiation hormesis. Hum. Exp. Toxicol., 19, 85–97.

17. Carmella, S. G., Chen, M., Yagi, H., Jerina, D. M. and Hecht, S. S. (2004): Analysis of phenanthrols in human urine by gas chromatography-mass spectrometry: potential use in carcinogen metabolite phenotyping. Cancer Epidem. Biomar. Prev., 13, 2167–2174.

18. Centers for Disease Control and Prevention (2005): Third National Report on Human Exposure to Environmental Chemicals. Atlanta, GA, USA, 2005.

19. Doll, R., Vessey, M. P., Beasley, R. W. R., Buckley, A. R., Fear, E., Fisher, R. E. W., Gammon, E. J., Gunn, W., Hughes, G. O., Lee, K. and Norman-Smith, B. (1972): Mortality in gasworkers – Final report of a prospective study. Brit. J. Ind. Med., 29, 394–406.

20. Environmental Protection Agency (1996): Proposed Guidelines for Carcinogen Risk Assessment. Office of Research and Development, Washington DC, USA, EPA/600/P-92/003C, Federal Register 61 (79), 17960–18011.

21. Greenland, S. (1995a): Dose-response and trend analysis in epidemiology: Alternatives to categorical analysis. Epidemiology, 6, 356–364.

22. Greenland, S. (1995b): Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology, 6, 450–454.

23. Grimmer, G., Dettbarn, G., Naujack, K. W. and Jacob, J. (1990): Excretion of hydroxy derivatives of polycyclic aromatic hydrocarbons of the masses 178, 202, 228 and 252 in the urine of coke and road workers. Int. J. Environ. An. Ch., 43, 177–186.

24. Hansen, M. H. and Kooperberg, C. (2002): Spline Adaptation in Extended Linear Models (with discussion). Stat. Sci., 17, 2–51.

25. Hastie, T. J. and Tibshirani, R. J. (1986): Generalized Additive Models. Stat. Sci., 1, 297–318.

26. Hastie, T. J. and Tibshirani, R. J. (1990): Generalized Additive Models. Volume 43, Monographs on Statistics and Applied Probability. London, UK: Chapman and Hall.

27. Hauptmann, M., Pohlabeln, H., Lubin, J. H., Jöckel, K. H., Ahrens, W., Brüske-Hohlfeld, I. and Wichmann, H. E. (2002): The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. Am. J. Ind. Med., 41, 89–97.

28. Hecht, S. S., Chen, M., Yagi, H., Jerina, D. M. and Carmella, S. G. (2003): r-1,t-2,3,c-4-Tetrahydroxy-1,2,3,4-tetrahydrophenanthrene in human urine: A potential biomarker for assessing polycyclic aromatic hydrocarbon metabolic activation. Cancer Epidem. Biomar. Prev., 12, 1501–1508.

29. Hecht, S. S., Chen, M., Yoder, A., Jensen, J., Hatsukami, D., Le, C. and Carmella, S. G. (2005): Longitudinal study of urinary phenanthrene metabolite ratios: Effect of smoking on the diol epoxide pathway. Cancer Epidem. Biomar. Prev., 14, 2969–2974.

30. Hemminki, K., Grzybowska, E., Chorazy, M., Twardowska-Saucha, K., Sroczynski, J. W., Putman, K. L., Randerath, K., Phillips, D. H., Hewer, A. and Santella, R. M. (1990): DNA adducts in human environmentally exposed to aromatic compounds in an industrial area of Poland. Carcinogenesis, 11, 1229–1231.

31. Hurvich, C. M. and Tsai, C. L. (1989): Regression and time series model selection in small samples. Biometrika, 36, 299–315.

32. International Agency for Research on Cancer (1983): Polycyclic aromatic compounds, chemicals environmental and experimental data. Volume 32, Monographs of the Evaluation of Carcinogenic Risks to Humans. Lyon, France.

33. International Agency for Research on Cancer (1984): Industrial exposures in aluminum production, coal gasification, coal production, and iron and steel founding. Volume 34, Monographs of the Evaluation of Carcinogenic Risks to Humans. Lyon, France.

34. International Agency for Research on Cancer (2006): Some non-heterocyclic polycyclic aromatic hydrocarbons and some related industrial exposures. Volume 92, Monographs of the Evaluation of Carcinogenic Risks to Humans. Lyon, France.

35. Jacob, J., Grimmer, G. and Dettbarn, G. (1999): Profile of urinary phenanthrene metabolites in smokers and non-smokers. Biomarkers, 4, 319–327.

36. Kim, Y. D., Lee, C. H., Nan, H. M., Kang, J. W. and Kim, H. (2003): Effects of genetic polymorphisms in metabolic enzymes on the relationships between 8-hydroxydeoxyguanosine levels in human leukocytes and urinary 1-hydroxypyrene and 2-naphthol concentrations. J. Occup. Health., 45, 160–167.

37. Kromhout, H. and Tongeren, M. van (2003): How important is personal exposure assessment in the epidemiology of air pollutants? Occup. Environ. Med., 60, 143–144.

38. Kuljukka-Rabb, T., Nylund, L., Vaaranrinta, R., Savela, K., Mutanen, P., Veidebaum, T., Sorsa, M., Rannug, A. and Peltonen, K. (2002): The effect of relevant genotypes on PAH exposure-related biomarkers. J. Expo. Anal. Env. Epid., 12, 81–91.

39. Li, C. S. and Hunt, D. (2004): Regression splines for threshold selection with application to a random effects logistic dose-response model. Comput. Stat. Data An., 46, 1–9.

40. Lin, Y. S., Kupper, L. L. and Rappaport, S. M. (2005): Air samples versus biomarkers for epidemiology. Occup. Environ. Med., 62, 750–760.

41. Lintelmann, J. and Angerer, J. (1999): PAH metabolites. In: Angerer, J. and Schaller, K. H. (Eds.): Analyses of Hazardous Substances in Biological Materials. Volume 6, Weinheim, Germany: Wiley-VCH, 163–187.

42. Loomis, D. and Kromhout, H. (2004): Exposure variability: Concepts and applications in occupational epidemiology. Am. J. Ind. Med., 45, 113–122.

43. Marczynski, B., Preuss, R., Mensing, T., Angerer, J., Seidel, A., El Mourabit, A., Wilhelm, M. and Brüning, T. (2005): Genotoxic risk assessment in white blood

cells of occupationally exposed workers before and after altering the PAH profile in the production material: comparison with PAH air and urinary metabolite levels. Int. Arch. Occ. Env. Hea., 78, 97–108.

44. Marczynski, B., Rihs, H. P., Rossbach, B., Holzer, J., Angerer, J., Scherenberg, M., Hoffmann, G., Brüning, T. and Wilhelm, M. (2002): Analysis of 8-oxo-7,8-dihydro-2'-deoxyguanosine and DNA strand breaks in white blood cells of occupationally exposed workers: Comparison with ambient monitoring, urinary metabolites and enzyme polymorphisms. Carcinogenesis, 23, 273–281.

45. McClean, M. D., Rinehart, R. D., Ngo, L., Eisen, E. A., Kelsey, K. T., Wiencke, J. K. and Herrick, R. F. (2004): Urinary 1-hydroxypyrene and polycyclic aromatic hydrocarbon exposure among asphalt paving workers. Ann. Occup. Hyg., 48, 565–578.

46. McCullagh, P. and Nelder, J. A. (1989): Generalized Linear Models. 2nd edition. London, UK: Chapman & Hall/CRC, Monographs on Statistics & Applied Probability.

47. Molinari, N., Daures, J. P. and Durand, J. F. (2001): Regression splines for threshold selection in survival data analysis. Stat. Med., 20, 237–247.

48. Moré, J. J. (1978): The Levenberg-Marquardt Algorithm: Implementation and Theory. In: Watson, G. A. (Ed.): Lecture Notes in Mathematics. Volume 630, New York, USA: Springer, 105–116.

49. Muggeo, V. (2003): Estimating regression models with unknown break-points. Stat. Med., 22, 3055–3071.

50. National Institute for Occupational Safety and Health (1994): Polynuclear aromatic hydrocarbons by HPLC. In: Cassinelli, M. E. and O'Connor, P. E. (Eds.): NIOSH Manual of Analytical Methods (NMAM). Washington DC, USA.

51. Nerurkar, P. V., Okinaka, L., Aoki, C., Seifried, A., Lum-Jones, A., Wilkens, L. R. and Le Marchand, L. (2000): CYP1A1, GSTM1, and GSTP1 genetic polymorphisms and urinary 1-hydroxypyrene excretion in non-occupationally exposed individuals. Cancer Epidem. Biomar. Prev., 9, 1119–1122.

52. Popp, J. A., Crouch, E. and McConnell, E. E. (2005): A Weight-of-Evidence analysis of the cancer dose-response characteristics of 2,3,7,8-Tetrachlorodibenzodioxin (TCDD). Toxicol. Sci., 89, 361–369.

53. Rihs, H. P., Pesch, B., Kappler, M., Rabstein, S., Rossbach, B., Angerer, J., Scherenberg, M., Adams, A., Wilhelm, M., Seidel, A. and Brüning, T. (2005): Occupational exposure to polycyclic aromatic hydrocarbons in German industries: Association between exogenous exposure and urinary metabolites and its modulation by enzyme polymorphisms. Toxicol. Lett., 157, 241–255.

54. Rosenberg, P. S., Katki, H., Swanson, C. A., Brown, L. M., Wacholder, S. and Hoover, R. N. (2003): Quantifying epidemiologic risk factors using non-parametric regression: Model selection remains the greatest challenge. Stat. Med., 22, 3369–3381.

55. Royston, P. and Altman, D. G. (1994): Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). Appl. Stat. – J. Roy. St. C, 43, 429–467.

56. Royston, P., Ambler, G. and Sauerbrei, W. (1999): The use of fractional polynomials to model continuous risk variables in epidemiology. Int. J. Epidemiol., 28, 964–974.

57. Royston, P. and Sauerbrei, W. (in press): Improving the robustness of fractional polynomials by preliminary covariate transformation: A pragmatic approach. Comput. Stat. Data An..

58. SAS Institute Inc. (2005): The GAM procedure. Cary, NC, USA, 2005, online publication: `http://support.sas.com/rnd/app/papers/gam.pdf`.

59. Scherer, G., Meger-Kossien, I., Angerer, J. and Knecht, U. (2001): Cotinine. In: Angerer, J. and Schaller, K. H. (Eds.): Analyses of Hazardous Substances in Biological Materials. Volume 7, Weinheim, Germany: Wiley-VCH, 171–189.

60. Schoenberg, I. J. (1946): Contributions to the problem of approximation of equidistant data by analytic functions, Part A: On the problem of smoothing or graduation, a first class of analytic approximation formulas. Q. Appl. Math., 4, 45–99, 112–142.

61. Schwarz, G. (1978): Estimating the dimension of a model. Ann. Stat., 6, 461–464.

62. Simpson, C. D., Wu, M. T., Christiani, D. C., Santella, R. M., Carmella, S. G. and Hecht, S. S. (2000): Determination of r-7,t-8,9,c-10-tetrahydroxy-7,8,9,10-tetrahydrobenzo[a]pyrene in human urine by gas chromatography/negative ion chemical ionization/mass spectrometry. Chem. Res. Toxicol., 13, 271–280.

63. Smyth, G. K. (2002): Nonlinear regression. In: El-Shaarawi, A. H. and Piegorsch, W. W. (Eds.): Encyclopedia of Environmetrics. Volume 3, Chichester, UK: John Wiley & Sons, Ltd., 1405–1411.

64. Steenland, K. and Deddens, A. D. (2004): A practical guide to dose-response analyses and risk assessment in occupational epidemiology. Epidemiology, 15, 63–70.

65. Stone, C. J. (1985): Additive regression and other nonparametric models. Ann. Stat., 13, 689–705.

66. Straif, K., Baan, R., Grosse, Y., Secretan, B., El Ghissassi, F. and Cogliano, V. (2005): Carcinogenicity of polycyclic aromatic hydrocarbons. Lancet, 6, 931–932.

67. Taussky, H. H. (1954): A microcolorimetric determination of creatine in urine by the Jaffe reaction. J. Biol. Chem., 208, 853–861.

68. Thayer, K. A., Melnick, R., Burns, K., Davis, D. and Huff, J. (2005): Fundamental flaws of hormesis for public health decisions. Environ. Health Persp., 113, 1271–1276.

69. Thurston, S. W., Eisen, E. A. and Schwartz, J. (2002): Smoothing in survival models: An application to workers exposed to metalworking fluids. Epidemiology, 13, 685–692.

70. Ulm, K. and Salanti, G. (2003): Estimation of the general threshold limit values for dust. Int. Arch. Occ. Env. Hea., 76, 233–240.

71. Umweltbundesamt (1998): Umwelt-Survey 1998 – PAK-Metaboliten im Urin [in German]. Dessau, Germany: Umweltbundesamt.

72. World Medical Association (1964): Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. "`http://www.wma.net/e/policy/b3.htm`".

73. Wu, M. T., Huang, S. L., Ho, C. K., Yeh, Y. F. and Christiani, D. C. (1998): Cytochrome P450 1A1 MspI polymorphism and urinary 1-hydroxypyrene concentrations in coke-oven workers. Cancer Epidem. Biomar. Prev., 7, 823–829.

74. Wu, M. T., Simpson, C. D., Christiani, D. C. and Hecht, S. S. (2002): Relationship of exposure to coke-oven emissions and urinary metabolites of benzo[a]pyrene and pyrene in coke-oven workers. Cancer Epidem. Biomar. Prev., 11, 311–314.

75. Zhao, L. P. and Kolonel, L. N. (1992): Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. Am. J. Epidemiol., 136, 464–474.

List of Abbreviations

AIC	Akaike information criterion
AICC	Akaike information criterion corrected
ANOVA	Analysis of variance
BGFA	Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin (Research Institute of Occupational Medicine of the German Social Accident Insurance)
BGIA	Berufsgenossenschaftliches Institut für Arbeitsschutz (Institute for Occupational Safety and Health of the German Social Accident Insurance)
BIC	Bayesian information criterion
CI	Confidence interval
crn	Creatinine
DF	Degrees of freedom
DTL	Double truncated linear
e. g.	Exempli gratia, for example
EPA	Environmental Protection Agency
GM	Geometric mean
GSD	Geometric standard deviation
g	Gram
HPLC	High performance liquid chromatography
i. e.	Id est, that is
LL	Log-likelihood
LQ	Limit of quantification
L	Liter
MAK	Maximale Arbeitsplatzkonzentration (Maximal concentration in the workplace)
OHPHE	Sum of 1-, 2-+9-, 3- and 4-OH-phenanthrene
PAH	Polycyclic aromatic hydrocarbon
PHE	Phenanthrene
QQ plot	Quantile-quantile plot
RSS	Residual sum of squares

List of Tables

List of Figures

A Double Truncated Power Functions form a Basis for Linear Splines

To show that the DTL functions defined in section 2.2.1.3 form a basis for the space of linear splines, it will be demonstrated, that each basis function of the truncated power basis can be constructed by linear combinations of the DTL functions.

It is

$$\begin{aligned}
\mathrm{TP}_{0,t,d=1}(x) &= 1 \\
\mathrm{TP}_{1,t,d=1}(x) &= (x - t_0) \\
\mathrm{TP}_{i,t,d=1}(x) &= (x - t_{i-1})_+ \\
&= I_{[t_{i-1},\,\infty)} \cdot (x - t_{i-1}) \quad \forall i = 2, \ldots, k+1 \; .
\end{aligned}$$

The functions $\mathrm{DTL}_{0,t}$, $\mathrm{TP}_{0,t,d=1}$ and $\mathrm{DTL}_{k+1,t}$, $\mathrm{TP}_{k+1,t,d=1}$ are already identical. For $i = 2, \ldots, k$ $\mathrm{DTL}_{i,t}$ can be written as follows

$$\begin{aligned}
\mathrm{DTL}_{i,t}(x) &= I_{[t_{i-1},\,t_i)} \cdot (x - t_{i-1}) + I_{[t_i,\,\infty)} \cdot (t_i - t_{i-1}) \\
&= I_{[t_{i-1},\,t_i)} \cdot (x - t_{i-1}) + I_{[t_i,\,\infty)]} \cdot (t_i - t_{i-1} + x - x) \\
&= I_{[t_{i-1},\,t_i)} \cdot (x - t_{i-1}) + I_{[t_i,\,\infty)]} \cdot \big((x - t_{i-1}) - (x - t_i)\big) \\
&= \big(I_{[t_{i-1},\,t_i)} + I_{[t_i,\,\infty)}\big) \cdot (x - t_{i-1}) - I_{[t_i,\,\infty)} \cdot (x - t_i) \\
&= I_{[t_{i-1},\,\infty)} \cdot (x - t_{i-1}) - I_{[t_i,\,\infty)} \cdot (x - t_i) \\
&= \mathrm{TP}_{i,t,d=1}(x) - \mathrm{TP}_{i+1,t,d=1}(x)
\end{aligned}$$

$$\Leftrightarrow \quad \mathrm{TP}_{i,t,d=1} = \mathrm{TP}_{i+1,t,d=1} + \mathrm{DTL}_{i,t} \; .$$

Consequently, the functions of the truncated power basis can be written in a recursive

manner as

$$\begin{aligned}
\mathrm{TP}_{k+1,t,d=1} &= \mathrm{DTL}_{k+1,t} \\
\mathrm{TP}_{k,t,d=1} &= \mathrm{DTL}_{k+1,t}(x) + \mathrm{DTL}_{k,t}(x) \\
&\vdots \\
\mathrm{TP}_{i,t,d=1} &= \sum_{j=i}^{k+1} \mathrm{DTL}_{j,t} \\
&\vdots \\
\mathrm{TP}_{1,t,d=1} &= \sum_{j=1}^{k+1} \mathrm{DTL}_{j,t} \\
\mathrm{TP}_{0,t,d=1} &= \mathrm{DTL}_{0,t} \ .
\end{aligned}$$

As the DTL functions and the truncated power basis consist of the same number of functions, the possibility to formulate the truncated power basis by linear combinations of the DTL basis is sufficient for the DTL functions to form a basis of the space of linear splines.

B Interpretation of Regression Parameters

Due to the highly skewed distributions, exposure to PHE in the workplace and urinary excretion of OHPHE were log-transformed for the analysis of the dose-response relationship. For the outcome OHPHE the transformation was performed using the natural logarithm, while for PHE the logarithm to the basis 2 was chosen. This leads to the following model (without regard for the confounders):

$$\ln(y) = \mu + \beta \log_2(x) + \varepsilon \,,$$

with y the excretion of OHPHE, x the exposure to PHE, μ the intercept, β the regression parameter of $\log(x)$ and ε the error term following a normal distribution.

For presentation, the parameter estimates of μ and β are transformed by the exponential function to allow a direct interpretation on the untransformed outcome:

$$\begin{aligned} y &= \exp(\mu + \beta \log_2(x) + \varepsilon) \\ &= \exp(\mu) \cdot \exp\big(\beta \log_2(x)\big) \cdot \exp(\varepsilon) \,. \end{aligned} \tag{1}$$

The first term $\exp(\mu)$ in (1) is the expectation of y under the restriction $\log_2(x) = 0$, i. e. $x = 1$. To interpret the second term, some considerations have to be made.

The expectation of the ratio of two function values can be written as

$$\begin{aligned} \frac{y_2}{y_1} &= \frac{\exp(\mu) \cdot \exp\big(\beta \log_2(x_2)\big)}{\exp(\mu) \cdot \exp\big(\beta \log_2(x_1)\big)} \\ &= \frac{\exp\big(\beta \log_2(x_2)\big)}{\exp\big(\beta \log_2(x_1)\big)} \\ &= \exp\big[\beta \cdot \big(\log_2(x_2) - \log_2(x_1)\big)\big] \\ &= \exp\big[\beta \cdot \log_2(x_2/x_1)\big] \,. \end{aligned} \tag{2}$$

If $\log_2(x_2/x_1) = 1$, that means if $x_2/x_1 = 2$, (2) reduces to $\frac{y_2}{y_1} = \exp(\beta)$. Consequently, $\exp(\beta)$ can be interpreted as the factor of alteration of the outcome y under a doubling of the predictor x.

Acknowledgments

An erster Stelle möchte ich mich bei Frau Dr. Beate Pesch, Chefin der Epidemiologie-Abteilung am BGFA in Bochum, und bei Herrn Prof. Dr. med. Thomas Brüning bedanken, die mir die Gelegenheit gegeben haben, die Daten des PAH Datensatzes statistisch zu bearbeiten. Beate hat mich tatkräftig in allen Ideen unterstützt und mich mit den nötigen biologischen und medizinischen Background-Informationen versorgt. Den Kollegen vom BGFA und vor allem der Epidemiologie-Abteilung möchte ich danken für wertvolle Diskussionen und eine schöne Zeit. Besonderer Dank gilt hierbei Sylvia Rabstein!

Ebenso möchte ich mich bedanken bei Herrn Prof. Dr. Marcus Neuhäuser und Herrn Prof. Dr. Karl-Heinz Jöckel für die Aufnahme als Doktorand am Institut für medizinische Informatik, Biometrie und Epidemiologie und die freundliche Betreuung dieser Arbeit.

Vielen Dank an Joachim Gerß und Kai Vogtländer, die sich aufgeopfert haben, um diese Arbeit auf Tipp- und sonstige Fehler hin zu überprüfen. Mille Merci à Annick Paisley qui a grandement contribué à l'améloriation de l'anglais de cette thèse.

Je suis particulièrement reconnaissant à Gilles Sonou, ancien chef d'Umanis CRO et maintenant Peter Holmes Clinical, qui était prêt à me donner du temps pour finir cette thèse. Merci pour ta confiance en moi! En même temps je remercie tous mes collègues d'Umanis et de Peter Holmes qui m'ont fait un accueil chaleureux en France. Vous avez toujours montré beaucoup de patience avec cet Allemand aux jeux bizarres. Un grand Merci spécial à Stéphanie Gautier et Stéphanie Seigle et toute l'équipe statistique.

Ein großer Dank gilt auch meinen Eltern, die mich mit viel Liebe erzogen haben, mir das Studium ermöglicht haben und immer für mich da waren, wenn ich sie brauchte. Vielen Dank auch an meine Geschwister und meine Oma. Ihr wart immer ein großer Rückhalt für mich.

Mein größter Dank geht an Sandrine. Vielen Dank für Deine Liebe, Deine Unterstützung, Deine enorme Geduld und Deinen Glauben an mich!

Curriculum vitae

Name	Martin Kappler
Date of Birth	25th May 1973 in Berlin
Address	Montreuil, France
Nationality	German
Marital Status	Maried

Employment

Since 12.2006	Senior Biostatistician, Peter Holmes Clinical, St. Denis, France
12.2005 – 11.2006	Senior Biostatistician, Umanis CRO, Levallois, France
04.2003 – 11.2005	Statistician and research associate at the Research Institute of Occupational Medicine of the German Social Accident Insurance (BGFA), Bochum, Germany
09.2002 – 01.2003	Teacher for mathematics and computer science at the Willy-Brandt school, Bottrop, Germany
08.2001 – 08.2002	Statistician and research associate at the Department of statistics at the University of Dortmund, Germany
07.2000 – 06.2001	Statistician and research associate at the Institute of biometry at the Medical University of Hannover, Germany
02.2000 – 06.2000	Statistician and research associate at the European Research Institute on Cancer and Immunology, Berlin, Germany

Education

04.2004 – 12.2007	PhD student at the Institute for Medical Informatics, Biometry and Epidemiology, Medical School of the University of Duisburg-Essen
12.1999	Master in Statistics (Diplom Statistiker), specialization Biometry
10.1993 – 12.1999	Study of Statistics, Faculty of Statistics, University of Dortmund
06.1992 – 09.1993	Civil service as ambulance man, Johanniter Unfall Hilfe, Dortmund
05.1992	General qualification for university entrance (Abitur), Sulz a. N.